DOCUMENT RESUME

ED 329 344 PS 019 445

AUTHOR Seppanen, Patricia S.; Love, John M.
TITLE Observational Study of Preschool Education and Care for Disadvantaged Children: Recommendations for Measuring Cognitive and Social-Emotional Outcomes among Chapter 1 Children.
INSTITUTION RMC Research Corp., Hampton, N.H.
SPONS AGENCY Department of Education, Washington, DC. Office of Planning, Budget, and Evaluation.
REPORT NO TAC-B-130
PUB DATE 15 Jul 90
CONTRACT LC-89098001
NOTE 103p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Cognitive Development; Compensatory Education; *Criteria; Day Care; *Disadvantaged Youth; Emotional Development; Guidelines; *Individual Development; Outcomes of Education; Preschool Education; Profiles; Research Design; Research Methodology; *Selection; Social Development; *Student Evaluation
IDENTIFIERS Education Consolidation Improvement Act Chapter 1

ABSTRACT

This paper presents recommendations about measures for assessing cognitive and social-emotional outcomes of children in Chapter 1 preschool and kindergarten programs. Section I explains the purpose and design of the study, giving special attention to the Chapter 1 substudy. Section II covers critical issues related to cognitive and social-emotional outcomes that will be measured as part of the substudy. Section III reviews basic considerations guiding the selection of measurement instruments and the supporting rationale. Section IV outlines the review process, summarizes criteria used in the review of instruments, and summarizes distinguishing characteristics of instruments that meet the criteria. Recommendations for instruments to be used in the study, and the rationale and description of necessary adaptations of one instrument, are included in Section V. Appendix A contains a summary of outcome measures and instruments used in large-scale studies in early childhood and recent state and local studies. Included in Appendix B is a preliminary screening of all candidate instruments. Appendix C includes profiles of instruments that meet preliminary criteria, while Appendix D includes a summary of responses to interviews with Chapter 1 program staff at the state and local levels regarding objectives, instructional approaches, and use of test instructments in Chapter 1 preschool programs. (RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TAC-B-130

PS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

X Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

ED329344

RMC
RESEARCH CORPORATION

Observational Study of Preschool Education and Care for Disadvantaged Children

Recommendations for Measuring Cognitive and Social-Emotional Outcomes Among Chapter 1 Children

BEST COPY AVAILABLE

PS 019445

RMC Research Corporation
400 Lafayette Road
Hampton, NH 03842

July 15, 1990

Observational Study of Preschool Education and Care for Disadvantaged Children

Recommendations for Measuring Cognitive and Social-Emotional Outcomes Among Chapter 1 Children

Patricia S. Seppanen
John M. Love

Submitted to:

U. S. Department of Education
Office of Planning, Budget and Evaluation
Washington, DC

Contract LC 89098001

RMC Research Corporation
400 Lafayette Road
Hampton, NH 03842

July 15, 1990

CONTENTS

		Page
Section I:	Purpose and Design of the Study	1
Section II:	Considerations in Selection of Measures	6
Section III:	Issues that Affect Measurement	8
	Validity	9
	Reliability	10
	Norm or Criterion Referenced	12
	Appropriate Norms	14
	Language Fairness	14
	Age Span	16
	Practical Considerations	16
Section IV:	Selection of Instruments	18
	Technical Characteristics	19
	Practical Considerations	19
	Effectiveness	20
	Characteristics of Instruments Meeting Criteria	20
	Candidate Cognitive Measures	21
	Candidate Social-Emotional Measures -- Rating Scales	23
	Candidate Child-focused Observation Systems	26
Section V:	Recommendations	27
	References	29
Appendix A:	Summary of Cognitive and Social-Emotional Measures Used in Similar Studies	
Appendix B:	Status of Cognitive and Social-Emotional Measures on Criteria for Initial Screening	
Appendix C:	Profiles of Instruments Meeting Preliminary Criteria	
Appendix D:	Chapter 1 Preschool Program Objectives, Instructional Approaches, and Testing Practices	

This paper presents recommendations of measures for assessing cognitive and social-emotional outcomes of children enrolled in Chapter 1 preschool and kindergarten programs. Section I explains the overall purpose and design of the study, with special attention to the Chapter 1 substudy. Section II discusses critical issues related to the cognitive and social-emotional outcomes that will be measured as part of the substudy. Section III reviews the basic considerations guiding the selection of measurement instruments, with the supporting rationale. Section IV outlines our review process, summarizes the criteria used in the review of possible instruments, and summarizes the distinguishing characteristics of instruments that meet these criteria. Recommendations for instruments to use in the study, including the rationale and description of necessary adaptations of one instrument are included in Section V. Appendix A contains a summary of outcome measures and instruments used in large-scale studies in the early childhood area and recent state/local studies. Included in Appendix B is a preliminary screening of all candidate instruments. Individual profiles of instruments that meet our preliminary criteria are included in Appendix C, while Appendix D includes a summary of responses to interviews with Chapter 1 program staff at the state and local levels regarding the objectives, instructional approaches, and current use of test instruments in Chapter 1 preschool programs.

Section I: Purpose and Design of the Study

Development Assistance Corporation of Dover, New Hampshire, in conjunction with subcontractors Abt Associates Inc. of Cambridge, Massachusetts, and RMC Research Corporation of Hampton, New Hampshire, is conducting an observation-based investigation of early childhood programs for the Office of Planning, Budget and Evaluation of the U.S. Department of Education.

This investigation is being conducted in coordination with other national and large-scale studies currently in place to update information from prior studies on the status and quality of child care and preschool programs. The primary purposes of the study are to:

- inform policymakers, early childhood educators, advocates, parents, and researchers about the relationships among program characteristics and indicators of program quality, and the impact on children of participating in early childhood programs; and
- develop a body of knowledge for dissemination to program administrators and child care providers that can influence the quality of programs.

This is a descriptive study of early childhood programs. In particular, this study is investigating variations in the quality of child care centers and preschool programs, in structural and environmental characteristics, in interactions between caregivers and children, and in the nature of children's activities in the centers and programs. The focus is on programs serving 4-year-old children who are economically disadvantaged. The study also includes a special substudy of Chapter 1 children, in which the influence of Chapter 1 preschool environments on children's cognitive and social-emotional development will be assessed.

The project is designed to allow us to address the following questions:

1. What is the range of young children's experience in early childhood programs?

   1a. How do children's experiences vary as a function of the characteristics of the site?

   1b. How do children's experiences vary as a function of the characteristics of the program?

   1c. How do children's experiences vary as a function of the characteristics of program staff?

2. What is the range of early childhood staff practices?

   2a. How does the concept of "developmentally appropriate practice" translate into curriculum, activities, instructional strategies, assessment, and discipline?

   2b. How are caregiver characteristics related to caregiver practice?

3. What are the relationships among different quality indicators?

Child outcome data that are collected as part of the Chapter 1 substudy will enable us to answer the following additional questions:

4. How are children's experiences and caregiver practice related to cognitive and social-emotional outcomes for children enrolled in Chapter 1 preschool programs?

   4a. What are the relationships between children's experiences and caregiver practice and outcomes when children's family background is taken into consideration?

   4b. How do these vary for different outcomes?

   4c. How stable are these outcomes for children from preschool to kindergarten?

5. How do the educational experiences of children enrolled in Chapter 1 programs change from preschool to kindergarten?

   5a. What discontinuities do children experience?

   5b. Are there relationships between discontinuities that children experience and outcomes in kindergarten?

   5c. How are they guided through the transition process?

6. For Chapter 1 preschool programs, can we begin to specify a range of acceptable quality variables based on the relationship between the quality indicators and outcomes for children enrolled in Chapter 1 preschool programs?

This study is being conducted in four low-income urban settings and one low-income rural setting, distributed among the four U.S. Census regions. An estimated 150 programs will be studied across the five sites. The programs include ones in public schools (including Chapter 1),

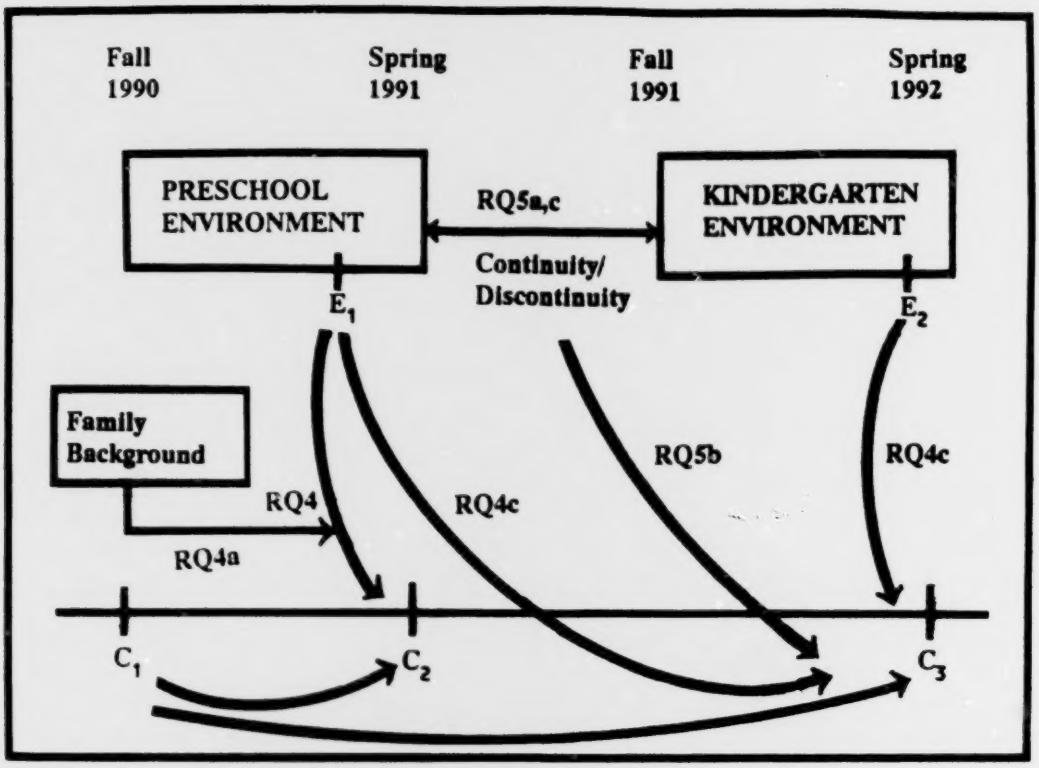
along with Head Start and other government and privately sponsored programs serving 4-year-old children who are disadvantaged, excluding family day care homes.

The assessments for the Chapter 1 substudy are being conducted on approximately 750 children enrolled in 25 Chapter 1 preschool programs (assuming two classrooms per program and 15 children per class). Individual cognitive and social-emotional assessments will be conducted in the fall and spring of the preschool year. Children will then be followed into their kindergarten year. Data on the kindergarten programs and the cognitive and social-emotional development of these children will again be collected during the spring of this school year. These data will allow us to chart the fall-spring-spring growth in these children and to relate that growth to features of their preschool and kindergarten programs. The statement of work prepared by the U.S. Department of Education for this project held out the possibility that these children would be followed into their elementary school years under another Department of Education contract to conduct a longitudinal study of children served through Chapter 1.

Figure 1 depicts the relationships that are being studied in the Chapter 1 substudy. This paper relates primarily to the child outcome measures that will be administered at points C_1, C_2, and C_3. Issues in measuring the preschool and kindergarten environments, the continuity/discontinuity between them, and family background characteristics are treated in separate papers.

The primary products of this study will be two reports -- one describing the findings and recommendations for both policy and research about the nature and quality of child care and early education programs for 4-year-old children who are disadvantaged, the other describing the findings and policy recommendations concerning Chapter 1 preschool programs.

Figure 1
Design of Chapter 1 Substudy to Examine the Effects of Children's Preschool and Kindergarten Environments on their Cognitive and Social-Emotional Development



Key:
- E_1 = Observations of preschool environment
- E_2 = Observations of kindergarten environment
- C_1 = Fall pretest measure of child outcomes
- C_2 = Spring of preschool measure of child outcomes
- C_3 = Spring of kindergarten measure of child outcomes
- RQ = Research question addressed by the arrow

Section II: Considerations in Selection of Measures

The review and selection of instruments for measuring child outcomes related to cognitive growth and social-emotional development brought up the issue of what we would measure related to these two broad constructs. In this process, we confronted a host of related issues. First, human behavior, particularly the behavior of young children, does not divide itself neatly into cognitive, social-emotional, motivational, personality, or physical development; the interaction among these domains is substantial (Aber, Molnar, & Phillips, 1986; Bradley & Caldwell, 1974; Goodwin & Driscoll, 1980; Katz & Jacobson, 1980).

A second, but related issue emerged when we began to review individual instruments: an instrument purporting to measure cognitive or social-emotional development may include tasks or questions that require responses involving a number of domains. While these additional domains (e.g., motivation, personality, physical development) are relevant to early childhood development, the cognitive and social-emotional domains are generally recognized by child development experts to be the important areas for a young child's development that can be directly influenced by participation in an early childhood program.

Third, past studies have found that certain cultural values in a child's home or community life may come into conflict with some of the behaviors valued by the public schools (Love, Wacker, & Meece, 1975; Raizen & Bobrow, 1974). Since we are interested in cognitive growth and social-emotional development as a function of participation in Chapter 1 pre-kindergarten and kindergarten programs, we suggest it is appropriate to limit our data collection to domains that are both valued and influenced by the public schools. At the same time, we are mindful of the need to select instruments that are fundamental enough in their assessment of cognitive and social-emotional development to avoid cultural bias.

Before beginning the selection process, we therefore sought answers to three questions related to the selection of outcome measures. First, what outcomes are of interest to Chapter 1 programs? In order to answer this question we made telephone calls to a number of local Chapter 1 directors to identify common approaches being used in Chapter 1 preschool and kindergarten classrooms. (Refer to Appendix D for a summary of responses.) Chapter 1 directors indicated that preschool programs tend to emphasize language enrichment and the development of basic skills or academic readiness skills. Preschool classrooms were typically described as using a developmental, activity, or experiential approach. Kindergarten classrooms were typically described as using an academic approach that is more teacher directed.

Second, we wanted to know the current thinking of child development experts regarding what outcomes are important developmentally for young children. To answer this question we reviewed articles by child development experts that discussed cognitive and social-emotional development. We found that although the "whole child" approach has been increasingly recognized by early childhood practitioners, child development theorists and researchers have been slow to respond with relevant theories and methods (Aber, et al., 1986). This is now beginning to change as scholars recognize more and more that cognitive development cannot be separated from social-emotional, motivational, personality or physical development (Block & Block, 1982; Cicchetti, Carlson, Braunwald, & Aber, 1986; Sroufe, 1979). These theorists increasingly recognize that advances and lags in one domain of development have implications for development in other domains, and that assessments of development are more sensitive and accurate when the interrelationships among domains are considered.

Researchers in academic achievement have been placing a growing emphasis on assessing children's adjustment to school and motivation to learn, as differentiated from their sheer intellectual capacity to learn (Aber et al., 1986). This development represents a shift away from

the more static measures of intellectual ability to the use of more dynamic assessments of classroom interactions, learning strategies, and motivational processes. Many researchers (Anderson & Messick, 1974; Scarr, 1981; Zigler & Trickett, 1978; Zigler & Seitz, 1980) have particularly stressed the importance of focusing on social or functional competence, which includes cognitive, social, and motivational components.

Finally, we asked what outcomes may be feasibly measured in the Chapter 1 substudy given the resources available. We concluded that issues related to cost feasibility, including time allocated for each testing situation, the training of examiners, and scoring, must be considered. These and other issues are discussed in Section IV of this paper as practical considerations.

Section III: Issues that Affect Measurement

In selecting or adapting a measurement instrument, it is important to ensure that it will actually measure what it is intended to measure, yield accurate scores, and be relatively easy to administer and score. These characteristics refer generally to the instrument's validity, reliability, and practical utility, respectively. We did not expect to find test instruments that met all relevant psychometric and use-related properties. For example, a very long test of reading readiness may yield more accurate scores than a shorter version of a similar test, but the longer test will take much more time to administer and perhaps require a more highly trained examiner.

In this section, we will discuss more of the technical issues related to validity, reliability, norming, and cultural fairness, as well as practical considerations such as compatibility with the curriculum approaches being used in Chapter 1 programs, test administration, scoring, and cost. The information presented here is to allow the reader and instrumentation panel members to make a more informed judgment about the adequacy of our instrument review process.

Validity

The validity of a measure is the extent to which it fulfills the purpose for which it was intended. A measure may be valid for one purpose but not for others; thus the question of validity always pertains to specific uses. As Cronbach observed, "One validates not a test, but an interpretation of data arising from a specified procedure" (Cronbach, 1971, p. 447). How to establish such validity remains a point of considerable debate among measurement experts, but three types of validity criteria are widely recognized:

- Content validity, permitting the test user to estimate how an individual child performs in the universe of situations the test is intended to represent;
- Criterion related validity permitting an inference to be made about the child's present or future performance on some other relevant test or task. (Predictive validity refers to inferences regarding future performance, while concurrent validity refers to inferences concerning performance observed or measured at approximately the same time as testing takes place.); and
- Construct validity, providing the basis for inference about children's relative standing on some theoretical construct (e.g., intelligence, cognitive ability, social competence, readiness) that is assumed to be a major determinant of their performance.

Although test publishers tend to emphasize content validity when documenting the quality of their instruments, some experts have argued that "for educational purposes, tests should have curriculum and instructional validity, i.e., they should be related to the content of curriculum and instruction" (Haney & Gelberg, 1980, p. 10). Such criticism has followed the national Follow Through evaluation (Bock, Stebbins & Proper, 1977), which compared 13 of the Follow Through models of early childhood education, using data based on a sample of over 20,000 Follow Through children over a four-year period. A major criticism of the findings was that "the outcome measures assess very few of the models' goals and strongly favor models that concentrate on teaching mechanical skills" (House, Glass, McLean, & Walker, 1978, p. 156). The Follow Through evaluation, in particular, taught us the power of local contextual variables. Models that

worked well in one community worked poorly in another -- unique features of the local setting had more effect on test scores than did the models, reminding us that "the most significant factors affecting educational achievement may be outside the control of public policies" (House et al., 1978, p. 156).

Given the purposes of the Chapter 1 substudy, we are particularly interested in test instruments that are validated for four types of inference: (1) how well they measure aspects of children's development, specifically cognitive and social-emotional development (construct validity); (2) how well they predict current and subsequent academic performance both in kindergarten and in later elementary, and perhaps even secondary, school (criterion-related validity); (3) how well they sample relevant aspects of cognitive and social-emotional development (content validity); and (4) how well the contents of the instruments match the contents of Chapter 1 preschool programs (content or instructional validity).

Although we expect to rely on the evidence presented by the test's publishers to establish the validity of an instrument, we will need to form our own judgments regarding content or instructional validity. This is often informed by empirical evidence from the use of instruments in large-scale studies. As an initial step in our review of individual standardized tests, we first investigated the local objectives and curricula of a sample of Chapter 1 preschool programs located across the United States. We then examined test items from instruments that met our other criteria, item by item, for representativeness.

Reliability

A measure is considered reliable if the scores it yields are consistent. Assessing the reliability of a test requires determining the precision of the measurement technique. A reliability estimate gives the expected consistency of scores for the measure.

Reliability is necessary, but not sufficient, for validity. To be reliable, a measurement must correlate reasonably well with itself. If it does not correlate with itself, it cannot correlate well with any external criterion either. However, a measure can be reliable without being valid. As compared with validity evidence, reliability evidence is relatively easy to obtain. For this reason, the reliability of many published instruments is documented in test manuals, while documentation of their validity is scant or unavailable. Reliability is secondary in importance to validity, and instruments accompanied only by information about their reliability cannot be considered adequate for use in the Chapter 1 substudy.

Three types of reliability are most commonly treated in the educational measurement literature:

- Internal consistency, referring to the extent to which all items or parts of an instrument measure the same thing;
- Alternate form reliability, meaning the comparative accuracy of results from equivalent forms of the same assessment instrument; and
- Stability, referring to the consistency of assessment results over time.

Some researchers have tried to rate the reliability of tests independently of test use, but this ignores the obvious point that reliability of assessment is more important for some uses than for others. For example, if the test is to be used to select children for ongoing participation in special services, then reliability matters more than if it being used as a periodic check on the progress of children. Given the purposes of the Chapter 1 substudy, we are most concerned with the stability of assessment results over time as the more rigorous reliability index. If we concluded that it was necessary to select subtests from an instrument, then separate reliability coefficients and details of the procedures used for obtaining them were included in our review.

Achieving internal consistency and stability with an instrument tends to be problematic when young children are involved (Brooks & Weintraub, 1976, p. 39; Walker, Bane, & Bryk, 1973, p.

26). Young children generally have shorter attention spans than older children -- at least for tasks that are not of their own choosing. As a result, it is important that assessment tasks for young children be kept short. But psychometrically, the shorter the test, the fewer items it has and, therefore, the lower its reliability. We can think of three ways to get around this problem: (1) keep the testing situation short (only 15 to 20 minutes per session to avoid problems of inattention and fatigue); (2) rely on instruments that are individually administered to help maintain children's interest; and (3) select tests that include tasks of intrinsic interest to children.

Norm or Criterion Referenced

Norm-referenced tests indicate relative performance by comparing the performance of individuals with that of a group of individuals taking the same test. The comparison with this "norm" group is typically made in terms of percentiles. Criterion-referenced tests provide information about performance on a specified criterion or set of criteria. The individual's performance is interpreted by comparison with pre-determined criteria, not with reference to a norm group (Scriven, 1980).

The choice of norm-referenced or criterion-referenced measures is not clear-cut and depends on a number of related factors. First, we must consider the nature of the data required, given the purpose of the Chapter 1 substudy. For example, if a test is to be used to select a certain number of children for a specialized remedial program, a norm-referenced test would be preferable. The children scoring near the bottom of the distribution would be selected. If, however, we wish to know the extent to which children have mastered certain objectives in structured instructional programs, a criterion-referenced test would be preferable because the purpose of the assessment is to determine the number of children who have achieved certain learning goals rather than to compare children with a national reference group.

Second, because Chapter 1 children may eventually be followed longitudinally (without a control group) through elementary or secondary school, the instruments we select must permit comparisons with outcome data obtained in future years. This would be most feasible with norm-referenced instruments.

Third, we must consider the pool of available instruments that adequately address the relevant dimensions of cognitive and social-emotional development in young children. Although reasonably good norm-referenced instruments exist in the cognitive domain, we know of no technically adequate normed instruments in the social-emotional area.

Finally, we expect some variation among Chapter 1 preschool and kindergarten programs. Since we are interested in how child outcomes are affected by generic attributes of program quality, we need child measures that are relatively independent of variations in particular program objectives or learning approaches. Several observers (Carver, 1974; Madaus, 1979; Popham, 1978) have directly criticized the widespread use of norm-referenced standardized tests for this reason. Precisely because of the way they are constructed, norm-referenced tests will theoretically be insensitive to the instructional effects of particular educational programs, so some critics advocate the use of criterion-referenced instruments. Other observers, however, suggest that both types of tests have a place in evaluation. The more curriculum-sensitive criterion-referenced tests can play an important role in program evaluation, while norm-referenced tests may be useful in comparisons of educational outcomes over time.

Thus, given the more short-term objectives of the Chapter 1 substudy and the more long-term possibility that children participating in the Chapter 1 substudy may eventually be followed longitudinally through their later school years, we recommend the use of norm-referenced instruments, if possible. In addition, because the measurement of children's social development remains a nagging problem, due largely to inadequate construct validation, we recommend the

selection of a criterion-referenced instrument in the social-emotional domain that is not related to particular program objectives or approaches. We recognized that this may necessitate the use of a classroom-based observation instrument and a rating scale that is completed by the classroom teacher rather than a test of social-emotional development.

Appropriate Norms

Many people believe that standardized tests are biased against particular subgroups of children, including minorities and children who are economically disadvantaged. We do know that a test may measure different functions when given to children who vary in sex, age, ethnicity, socioeconomic level, educational background, or other pertinent characteristics. Therefore, a test may demonstrate good stability and have high predictive validity when used with one group of children, but be much less stable and valid with other groups of children. Both validity and reliability coefficients should be accompanied by a full description of the samples used in obtaining them. One mistake commonly made in this connection is to assume that, because the norming sample includes some individuals who are like the individuals or group with whom a test is to be used, the test norms are therefore appropriate. Even if a norming group contains a ten percent sample of minority children, for example, the norms are not necessarily appropriate for use with minority children. Therefore, we examined the general characteristics of the overall norming sample. If the samples were not representative of the Chapter 1 children with whom the tests will be used (e.g., if they exclude disadvantaged or minority children), we dropped them from further consideration.

Language Fairness

A particular form of the more general problem of cultural bias in assessment is the issue of assessment of children whose native language is not standard English. This problem has most often been discussed with respect to Spanish-speaking children, but is obviously relevant to any

children who do not speak standard English as their native language. There is not space here to treat issues of bilingual assessment in any great detail. Nevertheless, a few basic points can be mentioned to highlight our decisionmaking process. First, it is important to distinguish linguistic or cultural differences among children from educational or learning differences, lest test performance be mistakenly interpreted as reflecting a learning "deficit." Second, even when assessments are carried out in children's native languages, the results of these assessments cannot be assumed to be equivalent to those of English-language assessments. Merely creating a literal translation in Spanish does not mean that the test's results with Spanish-speaking children will be equivalent to results from the English version with English-speaking children. Third, assessment of children who do not speak English as their native language must be viewed in light of the purposes of the Chapter 1 substudy. Using an English-language test with such children might be appropriate if the goal is to assess the program's success in teaching English to limited or non-English speaking children, but quite inappropriate if it is to measure children's general reading or math readiness. Finally, although the problem of cultural bias in written language tests is widely recognized, critics often overlook another problem: Assessments relying on pictorial representations may carry a burden of cultural dependency as great or even greater than those requiring verbal interaction (Anastasi, 1976).

We have established that the existence of non-English revisions of a particular test will be one of the criteria used in our preliminary review of candidate instruments. The option of excluding Spanish-speaking children from the sample was <u>not</u> considered (although it was necessary to exclude other non-English speaking children because appropriate instrumentation does not exist).

Age Span

Measures that span continuously the years four to six must encompass the range of characteristics representative of a period of particularly rapid development and must cover developmental transitions. By selecting measures that provide continuous coverage of a wide range of characteristics in the cognitive and social-emotional areas, it will be possible to establish a database on child outcomes that can lead to important insights on longitudinal relationships. Because one of the purposes of the Chapter 1 substudy is to measure change over time, and because we have not identified a control group, it is essential that we use comparable measures.

Practical Considerations

We must also address a number of practical considerations in the selection of measuring instruments, including compatibility with Chapter 1 programs, administration, scoring, and cost. These are discussed next.

Compatibility. In addition to establishing content or instructional validity, we must consider the current assessment instruments being used in local Chapter 1 programs. Selecting test instruments that are widely used in Chapter 1 programs will minimize unnecessary disruptions, offer school personnel an incentive to participate, and encourage parents to give their consent (e.g., arrangements might be made so that schools could use data from our administration for their own evaluation and reporting purposes). At the same time, our assessment activities must be compatible with Chapter 1 assessment activities in order to limit differences among children due to "test-wiseness" or familiarity and experience with test-taking procedures. Therefore, before we made our final recommendations of test instruments we considered information obtained from interviews conducted with a number of state and local public school personnel to identify the instruments that are most commonly used in Chapter 1 preschool and kindergarten programs. (See results of these interviews in Appendix D.)

Administration. The **amount of time** required for administration is an important practical consideration, especially with measures for young children. Test tasks can fail to hold a child's attention for a sufficient time period, thus increasing the difficulty of achieving measure reliability. Test reliability can theoretically be improved by adding comparable items, as we have noted; however, an important assumption is that the increased test length will not cause the children to become bored or inattentive. If they do, new sources of error, such as guessing, may be introduced and the incidence of missing data may increase. A test (or tests) that can be administered in less than 20 to 30 minutes or administered in two short testing sessions with a few hours or a day intervening between sessions, is a reasonable compromise between technical and practical considerations when testing young children.

A second major concern with regard to test administration is the **standardization of the stimulus situation**. We expect to use a number of junior professionals working in teams of two as examiners. Training and supervision of these examiners will help to ensure that each child receives an equivalent stimulus; but as an initial step, we will consider how clear and detailed the testing manuals are in specifying the testing procedures and instructions. Test format, content, testing conditions, and test-wiseness are common sources of "irrelevant difficulty" that can lead to less than accurate results. Therefore we reviewed individual tests to assess these "irrelevant difficulty" factors in order to minimize them.

The practical usefulness of a measure is further influenced by the **examiner training** required. Measures that can be administered by junior professionals who have been offered a reasonable amount of background and training are more convenient (as well as being less time consuming and costly) than measures, such as individually administered intelligence tests, that require extensive examiner training.

Scoring. Scoring procedures can also affect a test's usability. Objectivity and clarity of scoring procedures are particularly important qualities since we will be using junior professionals as examiners. As discussed above, we were also mindful of the amount of training required to achieve satisfactory administration and scoring by these personnel.

Cost. The Chapter 1 substudy calls for fall and spring assessments of 750 children enrolled in preschool programs and follow-up assessments during their kindergarten year. The time needed to complete these individual assessments (including time spent in set-up and transition in and out of the test situation) must conform to our budget allocation for examiners. Cost has implicitly been considered in many of the criteria discussed above (length of actual test situation, background and training of examiners, and scoring) so we did not consider it as a separate selection criterion per se.

Section IV: Selection of Instruments

Before beginning the review and selection process, we established a number of criteria by which all candidate measures would be judged. First, five basic criteria were used to conduct a preliminary screening of all the instruments in the cognitive and social-emotional areas that have been used in a large-scale national study in the early childhood area or in a recent state and local study. (Refer to Appendix A for a summary of the child outcome measures of cognitive and social-emotional development that were used in each study.) The five basic criteria include:

- Instrument must measure common Chapter 1 objectives in relevant domains of either cognitive skills or social-emotional development;
- Instrument must be appropriate for Chapter 1 children's ages and ability levels in preschool and/or kindergarten;
- Instrument, or selected sub-scales, can be administered in a reasonable time for young children (so that total testing time does not exceed 20 minutes);

- Examiners with backgrounds in child development and experience working with children may be trained in one day to administer and score the instrument;
- Instrument has been translated into Spanish, with evidence that all technical criteria are met.

Tables B.1, B.2, and B.3 in Appendix B summarize our preliminary screening of 29 instruments in the cognitive area and 26 instruments in the social-emotional area. Measures that met these five basic criteria were then examined further on the basis of the following additional criteria:

Technical Characteristics

- Content Validity: Consensus among child development experts that the instrument samples relevant dimensions of cognitive and social-emotional development;
- Instructional Validity: A majority of the test items overlap with the stated purposes and instructional approaches used in Chapter 1 preschool programs;
- Criterion Validity: Empirical evidence that establishes concurrent or predictive validity;
- Construct Validity: Empirical evidence that establishes that the instrument measures relevant aspects of children's development;
- Internal Consistency: Reliability coefficient (Cronbach's alpha or equivalent) of .65 or greater for test or subscales;
- Stability: Test-retest reliability coefficient of at least .65 for test or subscales;
- Appropriate Norms: The norming sample includes more than a token inclusion of children who are minorities and/or disadvantaged.

Practical Considerations

- Compatibility/Extent of Current Use: Instrument is being used in at least some Chapter 1 preschool programs;
- Method of Administration: Instructions clearly outline tasks for examiner or can be readily adapted for administration under study conditions;
- Scoring: Instrument requires a minimum degree of subjective scoring;
- Cultural Fairness: Empirical evidence is presented that indicates performance by children is not a function of subgroup membership.

Effectiveness

- Evidence of Effects Revealed: Evidence from past national or state/local studies indicates that the instrument has yielded credible data regarding child outcomes.

We decided to examine these important features in the second stage of our review process because if a measure failed to meet some of our more pragmatic criteria, there would be no need to consider it further, good psychometric qualities notwithstanding.

Characteristics of Instruments Meeting Criteria

The number of instruments coming close to meeting our five basic criteria was not large. There were four instruments in the cognitive area and four in the social-emotional area. Two additional instruments contain items related to both the cognitive and social-emotional domains. Detailed profiles of the following candidate instruments are included in Appendix C:

Cognitive Measures

- Brigance Preschool and K-1 Screen
- McCarthy Scales of Children's Abilities (MSCA)
- Peabody Picture Vocabulary Test -- Revised (PPVT-R)
- Preschool Inventory -- Revised (32 item version) (PSI)

Social-Emotional Measures

- California Preschool Social Competency Scale (CPSCS)
- Child Behavior Rating Scale (Version by RMC Research Corporation)
- Child Behavior Rating Scale (Version by Abt Associates)
- Howes Peer Play Scale

Measures of Both Cognitive and Social-Emotional Areas

- Battelle Developmental Inventory Screening Test (BATTELLE-S) or the Battelle Developmental Inventory (BDI)
- Bronson Social and Task Skill Profile

Our more indepth review of the technical characteristics, practical considerations, and evidence of effectiveness of these eight candidate measures indicated that the BATTELLE-S, Brigance Screen (both preschool and the K-1), and the McCarthy Scales of Children's Abilities (MSCA) should be dropped from further consideration because they do not in fact meet all of our basic criteria. The publishers of the BATTELLE-S and the MSCA indicate these tests are not available in languages other than English. Our evidence from studies using the MSCA indicates that Hispanic children were usually excluded from the sample or tested in English. One exception is the evaluation of Project Developmental Continuity (a Head Start demonstration program) that developed its own Spanish translation of several MSCA scales -- Verbal Memory - 1, Verbal Memory -2, Verbal Fluency, and Draw-A-Child (Love, Granville, & Smith, 1978) -- but data are available only on a relatively small sample of Hispanic children. The Brigance Screen is a criterion-referenced instrument, a test characteristic we decided not to consider in the cognitive area.

Discussed below are the characteristics of the two remaining candidate measures in the **cognitive area**.

Candidate Cognitive Measures

Description. The PPVT-R is an untimed test that typically takes 15 to 20 minutes to administer. Test items, arranged in order of increasing difficulty, consist of plates of four pictures. The PPVT-R may be used with children aged 2.5 and over. Children are shown a plate and asked to point to the picture that corresponds to the stimulus word pronounced by the examiner. A Spanish version of the PPVT-R, the Test de Vocabulario en Imagines Peabody (TVIP), has the same structure and standard score system. The aspect of cognitive ability measured by the PPVT-R and the TVIP is relatively narrow, restricted to receptive vocabulary. In addition, the

publishers have concluded that the PPVT-R is not a comprehensive measure of intelligence, but that it does help predict school success.

The Preschool Inventory -- Revised (PSI) is a 32-item test that is administered individually by an examiner. The test is untimed and takes approximately 15 minutes to administer. The PSI was developed originally to provide Head Start with a practical measure of preschool achievement and may be used with children aged 3 to 5 years. The test includes items of general knowledge, labeling, perception, and general concepts. The PSI uses a structured testing situation in which the examiner orally presents the test items. The child's response may be oral, pointing, or motor, as appropriate. A Spanish version of the PSI is available.

Technical characteristics. Both the PPVT-R and the PSI have strong psychometric characteristics. The PPVT-R has demonstrated adequate reliability and predictive validity with a variety of achievement and intelligence measures. Norms for the PPVT-R are based on a nationwide sample that was representative of the U.S. population according to the 1970 census. Minorities were included in the standardization process. Separate standardizations have been conducted on the TVIP with Spanish-speaking children in Mexico and Puerto Rico. Both combined and separate norms are available for the TVIP to interpret results.

The PSI has demonstrated its reliability and sensitivity to center- and home-based educational programs. Studies of validity and reliability are based on earlier versions of the PSI (containing 64 test items). The most recent version of the PSI (32 items) does not have national norms, although the evaluation of Project Giant Step (1989) in New York City does provide data from over 900 disadvantaged four-year-olds, including a number of Spanish-speaking children. In addition, reliability measures reported as part of Project Giant Step demonstrated the adequacy of the PSI in this area, and most of the national Head Start evaluations have reported strong internal consistency reliability for the 32-item PSI.

Practical considerations. Both the PPVT-R and the PSI are compatible with the focus and general approaches taken in many Chapter 1 preschool programs. However, one drawback of the PPVT-R is that it does not address other aspects of cognitive development that are relevant to child development and school readiness.

The PPVT-R and the PSI are both individually administered and take less than 20 minutes. Examiners for both instruments may be trained paraprofessionals.

Instructions for administration and scoring of each instrument are straightforward and clearly specified. Administration procedures for the PPVT-R require the child to respond to items by pointing to the picture that best illustrates the meaning of a stimulus word presented orally by the examiner. A score is obtained on the PPVT-R by subtracting errors from a total ceiling score and may be converted to a percentile rank, age equivalent score, or a standard score. All test items on the PSI are presented orally to a child and responses are scored as either correct or incorrect. A child's score on the PSI is the number of correct responses out of a maximum of 32.

Effectiveness. The PPVT-R and PSI have been used in a large number of national and large-scale studies involving young children. The PPVT-R has performed well consistently, although it has usually been used as part of a larger battery of tests measuring cognitive ability. The PSI has also consistently yielded significant results in terms of magnitude of change in child performance resulting from participation in early childhood programs.

The characteristics of the four remaining instruments in the **social-emotional** area are summarized below.

Candidate Social-Emotional Measures -- Rating Scales

Three of the instruments that were under consideration are paper-pencil rating scales completed by an adult who is well acquainted with the child. Two additional instruments,

recommended as candidate measures by the instrumentation panel, involve observations of individual children in their classroom settings. These observation instruments are described starting on page 26.

Description. The California Preschool Social Competency Scale (CPSCS) is a 30-item scale used to rate the interpersonal behavior of children between the ages of 2.6 and 5.6 years and the degree to which children assume social responsibility. The Child Behavior Rating Scale (CBRS-1) was created by RMC Research Corporation and used in an evaluation of home-based Head Start programs. The CBRS-1 is based on the Personal-Social and Adaptive Scales of the Battelle Developmental Inventory and measures interaction with adults, expression of affect, peer interaction, coping, social role, self-concept, and task mastery for children between the ages of 3 and 5. This 35-item instrument is typically completed by the child's teacher or home visitor, taking 10 to 20 minutes per child.

The second Child Behavior Rating Scale (CBRS-2) was adapted from the CBRS-1 by Abt Associates. The 34 items on this rating scale are based on coding categories from the CBRS-1 and the Bronson Social and Task Skill Profile Observation System. The CBRS-2 is designed to evaluate a child's social behavior with peers, with adults, and task behavior. The CBRS-2 takes a rater 10 to 15 minutes to complete per child.

Technical characteristics. As with most measures in the social-emotional area, limited information is available regarding the reliability and validity of the CPSCS, CBRS-1, or CBRS-2. The CPSCS has demonstrated adequate internal consistency and inter-rater reliability. The CPSCS is reported by its publisher to have face validity. No information is available regarding predictive validity. Measures of content, criterion, or construct validity are not available for the CBRS-1. As part of a pre-test in the Evaluation of the Home-based Option in Head Start, internal consistency was found to be very strong. Although test-retest reliability studies have not

been done, the fall-spring correlation for CBRS-2 ratings for children enrolled in Project Giant Step (a New York City program) was .67. Internal consistency (Cronbach's Alpha) was reported as .96 overall. There is weak validity information from Project Giant Step in that the CBRS-2 rating scale items related to task orientation/strategies (the more cognitively-oriented items) were reported to be more strongly correlated with Preschool Inventory scores than were items measuring adult and peer interaction (social items).

Norming information has not been developed for either the CBRS-1 or CBRS-2. The examiner manual of the CPSCS provides percentile norms for children by sex and age group. The norming sample is reported to include children of parents from "high and low occupational levels." A potential drawback regarding the CPSCS is that another reviewer (Mediax Associates, 1980) reports that this scale was normed primarily with middle-class children and that some test items may be culturally-biased.

Practical considerations. The CPSCS, CBRS-1, and CBRS-2 are very easy for a rater to complete. No special training is required; however, the adults completing each scale must be well acquainted with the child. Every item of the CPSCS is rated using a four-point scale arranged in order of increased competency. The CPSCS total raw score is the sum of the ratings for the 30 items.

The CBRS-1 uses a four-point scale in which the rater indicates how well the item describes the child. The CBRS-2 uses a five-point scale to indicate how frequently the child exhibits each behavior. A child's total score on the CBRS-1 or CBRS-2 is the mean rating across all the items on the particular scale. In addition, three subscores are available from the CBRS-2, one for child interactions, adult interactions, and task behavior.

Effectiveness. Each of these three social-emotional rating scales have yielded positive effects when used as part of a national or large-scale study in the early childhood area. The

CPSCS was used in 1972 as part of the original Head Start Longitudinal study. The CBRS-1 has been used in a recent national study, the Evaluation of the Home-Based Option in Head Start, but it yielded mixed results. One possible reason why significant program impacts did not emerge was attributed to a ceiling effect. Preliminary findings from Project Giant Step in New York City indicate that child ratings on the CBRS-2 did improve from fall to spring.

Candidate Child-focused Observation Systems

Description. The Bronson Social and Task Skill Profile provides a way of evaluating children's social behaviors, mastery behaviors, and their use of time, all within the classroom setting. The underlying hypothesis is that the concept of "executive" ability or skills can be applied to these three areas of performance. The term "executive skill" implies skill in recognizing the relevant cues, parameters, or rules of a situation; skill in predicting and planning possible sequences of events and outcomes of a situation; and skill in organizing and controlling both the self and the social or material "other" in a situation in order to effectively reach chosen goals.

The Howes Peer Play Scale is designed to measure social interactions with peers and friendships of young children in a group setting. Social interaction skills include ease of entry into play groups, play with peers, affective expressions, and other behaviors that lead to peer acceptance and popularity. Friendships are defined as stable, dyadic relationships marked by reciprocity and shared positive effects.

Technical characteristics. Limited information is available regarding the validity of the Bronson or the Howes. Both instruments require free choice time for observers to complete their observations. Highly structured and/or teacher-directed classrooms may limit opportunities for data collection. The Howes, in particular, is used by observers during free play periods.

Both instruments call for an observer to make multiple five- to ten-minute observations of individual children. As with classroom observation systems, both of these instruments are complex

and require careful training. Individual scores on the Howes are based on the proportion of time a child spends in each type of play situation. On the Bronson, a numerical rate or percent score is obtained for each of the behavior activities observed.

Effectiveness. Both the Howes and Bronson have yielded positive effects when used as part of a major early childhood study. The Howes, however, has been used primarily in pre-kindergarten settings; the Bronson has been used with older children as well.

Section V: Recommendations

The specification of our selection criteria, review process, and discussion of candidate instruments was presented to four experts in child development, early childhood education, and child care, who serve on the study's instrumentation panel. Panel members were in agreement that we should strive for as broad a picture as possible regarding outcomes for children. Critical outcomes for young children were seen by panel members as general readiness to learn rather than more content specific outcomes. Specific recommendations from panel members included:

- augment any paper-pencil assessment of social-emotional behavior with data based on direct observation of children within the classroom setting;
- assess a smaller random sample of children during the second and third testing episode in order to free-up any resources needed to carry out more labor-intensive classroom observations; and
- avoid measures in the social-emotional area that assess personality variables that vary little from one classroom setting to another.

After carefully considering the comments of the instrumentation panel and talking individually with Dr. Martha Bronson about possible adaptations to the Bronson Social and Task Skill Profile, we formulated the following recommendation. In the cognitive area, we plan to use the Preschool Inventory -- Revised (PSI) and its Spanish translation during the fall and spring of

the children's preschool year. We will use the Peabody Picture Vocabulary Test -- Revised (PPVT-R) and its Spanish version in the spring of the kindergarten year. The use of the PPVT-R is necessitated because of possible PSI ceiling effects with children over the age of five. Since the Chapter 1 substudy is concerned more with the stability of the process-outcome relationships rather than tracking developmental growth over time, shifting from the PSI to the PPVT-R for the kindergarten assessment will not be a problem. The PSI is particularly appropriate as a measure of preschool achievement or kindergarten readiness. The PPVT-R is technically very strong and offers the best alternative kindergarten measure. Because the PPVT-R may be used with older children, it will provide a baseline measure that any longitudinal study of these Chapter 1 children can build upon.

In the social-emotional area, we propose using the Child Behavior Rating Scale (CBRS-2), a teacher rating scale, during the fall and spring of the children's preschool year, and the spring of their kindergarten year. In addition, we will use an adaptation of the Bronson Social and Task Skill Profile (called the Bronson Social and Task Skills Profile, 1990 Revision) during the spring of the preschool year and again during the spring of the kindergarten year to augment the data in the social-emotional area from the CBRS-2.

Taken as a whole, the CBRS-2 and Bronson provide several important advantages for this study. The focus will be on in-classroom behavior of children rather than individual assessments of social behavior outside of the classroom setting. Information will be collected on a wide range of social and task behaviors so that we overcome the problem of considering social and cognitive variables in isolation from one another. Observations will be directed at the behavior of individual children rather than on groups or the classroom as a whole.

REFERENCES

Aber, J.L., Molnar, J. & Phillips, D. (1986). Action research in early education: A role for the philanthropic and research communities in the New York City initiative for four-year olds. New York: Unpublished manuscript prepared for the Foundation for Child Development.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: MacMillan.

Anderson, S., & Messick, S. (1974). Social competence in young children. Developmental Psychology, 10(2), 282-293.

Baumrind, D. (1970). Socialization and instrumental competence in young children. Young Children, 26(2), 104-119.

Bock, G., Stebbins, L.B., & Proper, E.C. (1977). Education as experimentation: A planned variation model, Volume IV-B, Effects of Follow Through Models. Cambridge, MA: Abt Associates Inc.

Block, J.H., and Block, J. (1980). The role of ego control and ego resiliency in the organization of development. In W.A. Collins (Ed.), Minnesota Symposia of Child Psychology, Vol. 13, 39-101. Hillsdale, NJ: Lawrence Erlbaum.

Brooks, J., & Weintraub, M. (1976). A history of infant intelligence testing. In M. Lewis (Ed.), Origins of intelligence in infancy and early childhood. New York: Plenum Press.

Carver, R. (1974). Two dimensions of tests: Psychometric and edumetric. American Psychologist, 29, 512-518.

Cicchetti, D., Carlson, V., Braunwald, K., & Aber, L. (1986). The Harvard Child Maltreatment Project: A context for research on the sequelae of child maltreatment. In R. Gelles & J. Lancaster (Eds.), Child abuse: A Biobehavioral perspective.

Goodwin, W.L., & Driscol, L.A. (1980). Handbook for measurement and evaluation in early childhood education. San Francisco: Jossey-Bass.

Haney, W., & Gelberg, W. (1980, December). Assessment in early childhood education. Unpublished manuscript. Cambridge, MA: The Huron Institute.

House, E.R., Glass, G.V., McLean, L.D., & Walker, D.F. (1978, May). No simple answer: Critique of the Follow Through evaluation. Harvard Educational Review, 48(2), 128-160.

Johnson, O.G. (1976). Tests and measurements in child development: Handbook I and II. San Francisco: Jossey-Bass.

Laosa L.M. (1977). Non-biased assessment of children's abilities: Historical antecedents and current issues. In T. Oakland (Ed.), Psychological and educational assessment of minority children. New York: Bruner/Mazel.

Love, J.M., Granville, A.C., & Smith, A.G. (1978, April). Final report of the PDC feasibility study, 1974-1977. Ypsilanti, MI: High/Scope Educational Research Foundation.

Love, J.M., Wacker, S., & Meece, J. (1975, June). A Process evaluation of Project Developmental Continuity. Interim Report II, Part B: Recommendations for measuring program impact. Ypsilanti, MI: High/Scope Educational Research Foundation.

Madaus, G. (1979, May). The sensitivity of measures of school effectiveness. Harvard Educational Review, 49, 207-230.

Mediax Associates, Inc. (1980, July). Readings in the social-emotional domain: A resource book for measures developers. (Contract No. HEW-105-77-1006) Westport, CT: Author.

Mitchell, J.V. (Ed.). (1985). The ninth mental measurements yearbook. Lincoln, NE: The University of Nebraska Press.

Popham, W.J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Raizen, S., & Bobrow, S.B., (1974). Design for a national evaluation of social competence in Head Start children. (R-1557-HEW). Santa Monica, CA: Rand Corporation.

Scarr, S. (1981). Testing for children: Assessment and the many determinants of intellectual competence. American Psychologist, 36, 1159-1166.

Scriven, M. (1985). Evaluation thesaurus. (2nd Ed.) Inverness, CA: Edgepress.

Sroufe, L.A. (1979) The coherence of individual development: Early care, attachment and subsequent developmental issues. American Psychologist, 34, 834-841.

Swertland, R., & Keyser, D. (Eds.). (1986). Tests (Second Edition.) Kansas city, MO: Test Corporation of America.

Walker, D.K., Bane, M., & Byrk, A. (1973). The quality of the Head Start planned variation data. Unpublished manuscript. Cambridge, MA: The Huron Institute.

Zigler, E.F., & Seitz, V. (1980). Early childhood intervention programs: A reanalysis. School Psychology, 9(4), 354-368.

Zigler, E.F., & Trickett, P.K. (1978). I.Q., social competence, and evaluation of early childhood intervention programs. American Psychologist, 33, 789-798.

APPENDIX A

SUMMARY OF COGNITIVE AND SOCIAL/EMOTIONAL MEASURES USED IN SIMILAR STUDIES

A. LARGE SCALE NATIONAL STUDIES (Publication date of final report/article)

A.1 Child Care Staffing Study (1989)
A.2 Evaluation of the Home-Based Option in Head Start (1988)
A.3 Child and Family Resource Program Evaluation (1982)
A.4 Project Developmental Continuity Evaluation (1982)
A.5 National Day Care Home Study (1981)
A.6 Home Start Follow-up Study (1979)
A.7 National Day Care Study (1979)
A.8 Head Start Transition Study (1978)
A.9 Evaluation of the Process of Mainstreaming Handicapped Children into Project Head Start (1978)
A.10 National Follow Through Evaluation (1977)
A.11 Home Start Demonstration Program Evaluation (1976)
A.12 Head Start Longitudinal Study (1972)
A.13 Head Start Planned Variation Study (1971)

B. RECENT STATE AND LOCAL STUDIES (Publication date of final report/article)

B.1 Evaluation of the quality of care and services in for-profit and non-profit centers/ Connecticut (1989)
B.2 At-Risk Preschool Program/Chicago, IL (1988-89)
B.3 Pre-K Program/Austin, TX (1985-88)
B.4 Pre-K Program/Wichita, KS (1982-87)
B.5 Preschool Kindergarten Longitudinal Study/Ohio (1986-ongoing)
B.6 Project Giant Step/New York City (1986-ongoing)
B.7 New Parents as Teachers/Missouri (1983-84; published in 1985)
B.8 All-day Kindergarten/NYC (1983)
B.9 Bermuda Child Care Study (data collected in 1980; published in 1987)
B.10 Daycare programs for disadvantaged/Bermuda (1980)
B.11 Proprietary Day Care Centers/North Carolina (data collected in early 1980s; published in 1986)
B.12 Brookline Early Education Project (BEEP)/Brookline, MA (late 1970s)
B.13 Pre-K Program/New York State (1979)
B.14 Carolina Abecedarian Project (1970s)
B.15 Family Development Research Program/Syracuse University (1970s)
B.16 High/Scope Preschool Curriculum Comparison Study (late 1960s, early 1970s and later follow-up)

C. OTHER DATA COLLECTION EFFORTS

C.1 National Longitudinal Surveys of Youth/Child Assessments (1986)

SUMMARY OF COGNITIVE AND SOCIAL-EMOTIONAL OUTCOME MEASURES USED IN SIMILAR STUDIES
A. LARGE SCALE NATIONAL STUDIES

Name of Program	Reference	Outcome Measures	Comments
A.1 Child Care Staffing Study	Whitebook, et al. (1989, Nov.) Who cares? Child care teachers and the quality of care in America, Young Children, 45(1).	Security/attachment sociability: ■ Waters and Deane, Attachment Q-Set ■ Howes Peer Play Scale Communication skills: ■ Feagans and Farron Adaptive Language Inventory Language development: ■ Peabody Picture Vocab. Test - Revised (PPVT-R)	■ Child assessments conducted in 1 of 5 cities visited (Atlanta); no indication if limited English speaking children were assessed.
A.2 Home-Based Option in Head Start	Meleen, P., Love, J., & Nauta, M. (1988). Final report, Vol. I: Technical report. Study of the home-based option in Head Start. Hampton, NH: RMC Research Corp.	Cognitive: ■ Preschool Inventory (PSI Version R with 32 items) Social/Health: ■ Interview using Head Start Meas. Battery (social scale) Social: ■ Home visitor or teacher ratings using Child Behavior Rating Scale (CBRS)	■ Limited English speaking children were excluded from the sample; all testing was done in English. ■ CBRS was adapted by RMC Research from items on the Battelle Developmental Inventory.
A.3 The Child and Family Resource Program	Nauta, M. et al. (1982). The effects of a social program: Final report of the Child and Family Resource Program's infant/toddler component. Cambridge, MA: Abt Associates.	Child development and achievement: ■ Preschool Inventory (PSI Version R with 32 items) ■ High/Scope Pupil Observation Checklist (POCL) ■ Schaefer Behavior Inventory	■ Final report states that small groups of Hispanic families and families of other ethnic origins were excluded from quantitative analyses; reasons not given.

Name of Program	Reference	Outcome Measures	Comments
A.4 Project Developmental Continuity Evaluation	Bond, J.T. et al. (1982). Project developmental continuity evaluation final report. Ypsilanti, MI: High/Scope Educational Research Foundation.	Specific academic achievement: ■ Peabody Individual Achievement Test - Reading - Math ■ Metropolitan Ach. Test - Reading General academic skill/aptitude: ■ WPPSI - modified block design test ■ Bilingual Syntax Measure - English/Spanish versions administered to Spanish-dominant children ■ McCarthy Scales of Children's Ability - verbal fluency - verbal memory, Part I and II - draw-a-child Social Development/Adjustment: ■ Preschool Interpersonal Problem-Solving Test (PIPS) (adapted) ■ High/Scope Pupil Observation Checklist completed by teachers (to measure sociability) ■ PDC Child Rating Scale completed by teachers to measure independence (two items) ■ PDC Child Rating Scale completed by teachers to measure social adjustment (six items)	■ Spanish speaking children were tested in their native language; scores on Spanish version of the child battery did not appear to be equivalent to the English version, so were excluded from overall analyses. A separate, exploratory analysis was conducted of bilingual program effects.

Name of Program	Reference	Outcome Measures	Comments
A.4 Project Developmental Continuity Evaluation (Continued)		Attitude toward teacher/school: ■ PDC Child Interview (8 questions to measure attitude toward school) ■ PDC Parent Interview (to measure child's attitude toward school) ■ School attendance Learning attitude/style: ■ High/Scope Pupil Obs. Checklist (to measure task orientation) ■ PDC Child Interview/Scale 2 (3 questions to measure interest in reading) ■ PDC Child Rating Scale completed by teachers (7 questions to measure learning orientation)	
A.5 National Day Care Home Study	Divine-Hawkins, P. (1981). Family day care in the U.S.: National day care home study, Volume I. DHHS Publication 80-30287. Washington, DC: DHHS.	Caregiver and child behavior: ■ Carew/SRI Adult Behavior Codes and Child Codes	■ Data collected in natural situation within setting and in experimentally structured situation; observation data systems developed specifically for this study. ■ Hispanic caregivers represented in sample; collection of data in language other than English not indicated.

Name of Program	Reference	Outcome Measures	Comments
A.6 Home Start Follow-up Study	Nauta, M.J. et al. (1979). Home Start follow-up study: A study of long-term impact of Home Start on program participants. Cambridge, MA: Abt Associates.	Academic achievement: ■ Peabody Indiv. Achievement Test (Math & Reading Recognition Subtests) School adjustment: ■ Purdue Social Attitude Scale ■ Stephens-Delys Reinforcement Contingency Interview ■ Preschool Interpersonal Problem Solving Test ■ Parent Interview	■ No indication if limited English speaking children were assessed.
A.7 National Day Care Study	Ruopp, R.R., Travers, J., Glantz, F., & Coelen, C. (1979). Children at the center. Summary findings and their implications. Cambridge, MA: Abt Associates.	Knowledge: ■ Caldwell Preschool Inventory (PSI) Receptive vocabulary: ■ Peabody Picture Vocabulary Test (PPVT) Caregiver and child behavior: ■ SRI Preschool Obs. Instrument. Adult-Focus Instrument (AFI) and Child-Focus Instrument (CFI). ■ Child Dev. Assoc. Checklist (CDA) ■ Daycare Forces Inventory (DCFI)	■ Analyses focused on children's fall to spring gains calculated to avoid certain technical problems posed by simple difference scores. ■ No indication if limited English speaking children were assessed. ■ Adjusted gain scores almost completely independent of racial, socio-economic and other background characteristics.

Name of Program	Reference	Outcome Measures	Comments
A.8 Head Start Transition Study	Royster, E.C., and Larson, J.C. (1978). Executive summary of Head Start graduates and their peers. Cambridge, MA: Abt Associates, Inc.	Academic readiness: ■ Wide Range Ach. Test Attitudes: ■ Values Inventory for Children Friendship: ■ Informal questioning Social: ■ Teacher ratings on the Schaefer Classroom Beh. Inventory ■ Teacher ratings on the Beller Rating Scales Test orientation and sociability: ■ Teacher and examiner ratings on the Child's Test Orientation and Sociability	■ Post-testing only. ■ Sites enrolling primarily Hispanic children excluded from study.
A.9 Mainstreaming Handicapped Children into Head Start	Vogel, R. and Rader, J. (1978). Evaluation of the process of mainstreaming handicapped children into project Head Start. Phase II final report. Silver Spring, MD: Applied Management Science, Inc.	Physical self-help, social/emotional, academic development, communication: ■ Alpern-Boll Developmental Profile Classroom social behaviors and social integration: ■ California Preschool Social Competency Scale ■ Prescott-SRI Child Observation System	Sample consisted of children with handicapping conditions. Hispanic children were included in sample.

Name of Program	Reference	Outcome Measures	Comments
A.10 National Follow Through Evaluation	Stebbins, L.B. et al. (1977). Education as experimentation: A planned variation model IV.A, An evaluation of Follow Through. Cambridge, MA: Abt Associates.	Achievement: ■ Metropolitan Achievement Test Non-verbal problem solving: ■ Raven's Coloured Progressive Matrices (modified) Self-esteem: ■ Coopersmith's Self-Esteem Inventory Locus of Control: ■ Intellectual Achievement Responsibility Scale (modified)	■ Full battery used at end of third grade; MAT alone used at end of each preceding year. ■ No indication if limited English speaking children were assessed.
A.11 Home Start Demonstration Program (1972-75)	Love, J.M., Nauta, M.J., et al. (1976). National Home Start evaluation: Final report -- findings and implications. Ypsilanti, MI: High/Scope Educational Research Foundation.	School readiness and physical development: ■ Preschool Inventory Experimental Revision containing 32 items (PSI) ■ Denver Developmental Screening Test (DDST) ■ Child 8-Block Task Social-emotional development: ■ Schaefer Behavior Inventory (SBI) ■ Pupil Obs. Checklist (POCL)	■ Other child measures regarding: physical development, nutrition and medical care. ■ Non-English speaking families were excluded from the evaluation activities.

Name of Program	Reference	Outcome Measures	Comments
A.12 Head Start Longitudinal Study (1968-72)	Educational Testing Service (1968). Disadvantaged children and their first school experiences. ETS-OEO longitudinal study. Theoretical considerations and measures strategies. Princeton, NJ: Author. Emmerich, W. (1971). Structure and development of personal-social behaviors in preschool settings. ETS longitudinal study. Princeton, NJ: ETS.	Reasoning and Analytic: ■ Block Design: WPPSI and WISC ■ ETS Logical Reasoning Tests ■ Hess and Shipman 8-Block Sorting Task ■ Picture Block Test ■ Picture Completion: WPPSI and WISC ■ Portable Rod-and-Frame Test Attention, Learning, and Memory: ■ Animal House: WPPSI ■ Fixation Time ■ Form Memory ■ Fruit-Distraction Test ■ Relevant Redundant Cue Concept Task ■ Stanford Memory Test Attitudes, Interests: ■ Brown IDS Self-Concept Referents Test ■ Northeastern University Interest Inventory (adapted) ■ Social Schemata Controlling Mechanisms: ■ I-E Scale (Locus of Control) ■ Kreitler Cognitive Orientation ■ Matching Familiar Figures Test ■ Mischel Technique ■ Modified Hertzig Procedure ■ Motor Inhibition Test ■ Risk-Taking Tasks ■ Siegel Conceptual Style Sorting Task General Knowledge: ■ Preschool Inventory (64 items) ■ TAMA General knowledge	■ All or part of 81 measures of cognitive and perceptual development, personal and social development, and physical health and nutritional status were proposed. Only measures related to cognitive and social/emotional development and targeted for use with children aged three to six are listed here. ■ Children from families speaking a foreign language and those with severe physical handicaps excluded from sample.

Name of Program	Reference	Outcome Measures	Comments
A.12 Head Start Longitudinal Study (1968-72) (continued)		Perception: ■ Analysis of Visually Perceived Forms ■ Auditory Discrimination Test ■ Children's Auditory Discrimination Inventory ■ Developmental Test of Visual-Motor Integration ■ Johns Hopkins Perceptual Test ■ Seguin form Board ■ Synthesis of Visually Perceived Forms ■ Visual Perception Inventory: Position in Space Subtest Piagetian: ■ Conception of Natural Events ■ Conservation of Number ■ ETS Spatial Egocentrism Task ■ ETS Enumeration ■ Physical Identity and Sex Role Constancy Tasks ■ Spontaneous Correspondence Verbal: ■ ETS Communication Skills V-5 ■ ETS Matching Pictures Comprehension Task ■ ETS Story Sequence Task, Parts I and II ■ Harrison-Stroud Reading Readiness Profiles, Test 6 ■ Harvard Story Completion Test ■ Illinois Test of Psycho-Linguistic Abilities, Auditory-Vocal Automatic Subtest ■ Massad Mimicry Test ■ Metropolitan Readiness Tests	

Name of Program	Reference	Outcome Measures	Comments
A.12 Head Start Longitudinal Study (1968-72) (continued)		■ Peabody Picture Vocabulary Test (adapted) ■ TAMA, Tell-a-Story Task General Personality: ■ Classroom ratings to monitor each child's personal-social development using bipolar scales and unipolar personality characteristics (Emmerich)	
	Lee, V.E., et al. (1989). Are Head Start effects sustained? Unpublished manuscript.	Verbal Achievement: ■ Cooperative Primary Test Perception: ■ Children's Embedded Figures Test ■ Raven's Colored Progressive Matrices Social Competency: ■ Teacher ratings on California Preschool Competency Scale ■ Teacher ratings on Schaefer Classroom Behavior Inventory	■ Outcome measures listed here are from Head Start Longitudinal Study and were used in re-analysis.

Name of Program	Reference	Outcome Measures	Comments
A.13 Head Start Planned Variation Study	Klein, J. and Datta, L. (1971, November). Head Start planned variation study. Washington, D.C.: U.S. Department of HEW/Office of Child Development.	Receptive language: ■ Peabody Picture Vocabulary Test (PPVT) Knowledge of basic concepts: ■ Caldwell Preschool Inventory Baseline on letters/numbers: ■ Wide Range Achievement Test Self-concept: ■ Brown Self-Concept Test Child behavior: ■ Teacher ratings using Schaefer Behavior Inventory Interaction between child/mother: ■ Hess-Shipman 8-Block Sort Task Counting/matching conserving mass: ■ Enumeration Active/imaginative language: ■ Illinois Test of Psycholinguistic Ability (verbal expression subtest only) Capacity to inhibit movement: ■ Maccoby Motor Inhibition Test Object categorizing: ■ Sigel Object Categorizing Test	■ Information regarding linguistic backgrounds of children not presented.

B. RECENT STATE AND LOCAL STUDIES

Name of Program	Reference	Outcome Measures	Comments
B.1 Evaluation of the quality of care and services in for-profit and non-profit centers/ Connecticut (1989)	Kagan, S.L., & Newton, J.W. (1989, November). For-profit and non-profit child care: Similarities and differences, Young Children, 45(1), 4-10.	Domains not specified: ■ Trained observers used a modified version of the Child Development Associate (CDA) Checklist, and ■ A Child Behavior Scale (created for this study)	■ Centers visited included those with staff who were Black or other racial minorities; no information presented regarding observation of limited English speaking children.
B.2 At-Risk Preschool Program Chicago, IL (1988-89)	Jeanne B. Borger Department of Res. & Eval. 5W(n) Chicago Public Schools 1819 W. Pershing Road Chicago, IL 60609	Language and readiness: ■ Chicago Early Assessment (EARLY) Receptive language: ■ Peabody Picture Vocab. Test - Revised (PPVT-R) Social-emotional: ■ Observation; Parent Interview Domain not specified: ■ Criterion referenced tests Expressive/receptive language: ■ Preschool Language Assessment Scale (PRELA) used only with LEP and Non-English speakers in kindergarten	■ All screening done by bilingual staff in language with which a child was most comfortable (22% of children were Hispanic).
B.3 Pre-K Program (serving low-achieving and LEP students in Chapter 1 schools) Austin, TX (1985-88)	Dr. Catherine Christner Office of Research and Evaluation Austin Indep. School District 6100 Guadalupe, Box 79 Austin, TX 78752	Receptive Vocabulary: ■ Peabody Picture Vocab. Test-Revised (PPVT-R) ■ Test de Vocabulario en Imagenes Peabody (TVIP)	■ PPVT-R given to all students. ■ TVIP given to children who are Spanish monolingual; TVIP has same structure and standard score system as PPVT-R.

	Name of Program	Reference	Outcome Measures	Comments
B.4	Pre-K Program (Chapter 1 funding) Wichita, KS (1982-87)	Carolyn Max, Program Evaluation Jacqualyn L. Farha Pupil Evaluation and Testing Wichita Public Schools	Domains not specified: ■ Cooperative Preschool Inventory ■ DIAL-R (motor section) ■ Illinois Test of Basic Skills (ITBS) (used in 2nd/3rd yr. follow-up)	■ No adaptations reported for possible limited English proficient children (Asian, Hispanic and Native American).
B.5	Pre-school Kindergarten Longitudinal Study (Ohio) (1986 - ongoing)	Sheehan, R., Cryan, J., & Wiechel, J. (1989). Factors contributing to success in elementary schools: Results of a longitudinal study in process. San Francisco, AERA.	Achievement data: ■ Metropolitan Readiness, Version 5, Level 2 ■ Metropolitan Achievement (MAT6) ■ California Achievement (CAT) School behavior: ■ Hahneman Elementary School Behavior Rating Scale (HESB)	■ No indication if limited English speaking children were assessed.
B.6	Project Giant Step/ New York City (1986-ongoing)	Abt Associates. (1988, October). Evaluation of Project Giant Step; Technical Progress Report #4. Cambridge, MA: Abt Associates.	Domains not specified: ■ Preschool Inventory - 32 items (PSI) ■ Child Behavior Rating Scale ■ Bronson Executive Skill Profile	■ Adapted from CBRS; used in Home-Based Study and Bronson Executive Skill Profile ■ Assessment materials translated for use with children speaking Spanish or Yiddish.

Name of Program	Reference	Outcome Measures	Comments
B.7 New Parents as Teachers/Missouri (1983-84 report published in 1985)	Pfannenstiel, J.C. & Seltzer, D.A. (1985). Evaluation report: New parents as teachers project. Overland Park, KS: Research and Training Associates.	Social: ■ Battelle Developmental Inventory (BDI) Language: ■ Zimmerman Preschool Language Scale (PLS) Cognitive: ■ Kaufman Assessment Battery for Children (KABC) Parent Knowledge: ■ Parent Knowledge Survey Instrument Hearing: ■ Retrospective interview with parents and informal whisper test	■ No indication that limited English proficient children included in sample. ■ Parents assessed their child's social development using selected dimensions of the BDI. ■ Parent survey developed by Pfannenstiel and Seltzer.
B.8 All-day Kindergarten/ New York City (1983)	Jarvis, C.H., Molnar, J.M., and Collins, C. (1985). Urban education: can all-day kindergarten make a difference? Paper presented at the annual meeting of the American Psychological Association, Los Angeles, CA.	School-readiness: ■ Brigance K and 1 screening English language proficiency: ■ Language Assessment Battery	■ Sampled children included 50% who came from non-English speaking homes representing 24 linguistic-cultural groups. Primary non-English native language represented in sample was Spanish. ■ Brigance and LAB were routinely administered by NYC schools to determine eligibility for special programs.

	Name of Program	Reference	Outcome Measures	Comments
B.9	Bermuda Child Care Study (data collected in 1980)	Phillips, D. et al. (1987). Child-care quality and children's social development. Developmental Psychology, 23(4), 537-543.	Social development: ■ Classroom Behavior Inventory (preschool form) -- Schaefer and Edgerton Social adjustment: ■ Preschool Behavior Questionnaire (Behar, 1977)	■ Completed by two caregivers on each child and child's parent. ■ No indication any children were limited English proficient.
B.10	Daycare programs for disadvantaged children (data collected in 1980)	McCartney, K. et al. (1985). Day care as intervention: Comparisons of varying quality programs. Journal of Applied Developmental Psychology, 6, 247-260.	Domains not specified: ■ Peabody Picture Vocab. Test Revised (PPVT-R) ■ Preschool Language Assessment Instrument ■ Caregiver ratings on the Adaptive Language Inventory and research team ratings on a communication task Social skill: ■ Preschool version of the Classroom Behavior Inventory ■ Preschool Behavior Questionnaire	Communication task admin. to subset of children at each center. Two caregivers and parents interviewed.
B.11	Proprietary Day Care Centers/North Carolina (date article published: 1986; data collected in early 1980s)	Bjorkman, S. et al. (1986). Environmental ratings and children's behavior: Implications for the assessment of day care quality. American Journal of Orthopsychiatry, 56(2), 271-277.	Social behavior: ■ Social Observation Code developed by Poteat and Saudargas	■ No indication any children were limited English proficient.

Name of Program	Reference	Outcome Measures	Comments
B.12 Brookline Early Education Project (BEEP)/Brookline MA (late 1970s)	Pierson, D.E., Walker, D.K., and Tivnan, T. (1984, April). A school-based program from infancy to kindergarten for children and their parents. The Personnel and Guidance Journal, 62(8), 448-455. Pierson, D.E. et al. (1983, April). The impact of early education measured by classroom observations and teacher ratings of children in kindergarten. Evaluation Review, 7(2), 191-216.	Social and beh. performance: ■ Executive Skill Profile (Bronson, 1975; 1978) Social, pre-academic, motor, work behavior: ■ Teacher ratings using the Kindergarten Performance Profile (developed by Brookline staff)	■ Testing on outcome measures occurred during kindergarten. ■ Ten periodic assessments of children's physical, sensory, neurological and developmental status completed before kindergarten enrollment. Parents observed exam and discussed results, follow-up. ■ No indication if instruments were adapted for children whose first language in the home was not English.
B.13 Pre-K Program/ New York State (1979)	Irving, D.J. et al. Parent involvement affects children's cognitive growth. Cited in Collins, R.C. et al. (1982). The impact of Head Start on children's cognitive development. Washington, D.C.: Department of Health/Human Services.	General reasoning: ■ Walker Readiness Test for Disadvantaged Children School related knowledge/skills: ■ Cooperative Preschool Inventory Verbal concepts: ■ Peabody Picture Vocabulary Test	■ No information as to whether adjustments were made in assessment of limited English proficient children.
B.14 The Carolina Abecedarian Project (serving children from infancy to 6 1/2 years during 1970s)	Ramey, C.T. & Campbell, F.A. in Gallager, J.J. and Ramey, C.T., The malleability of children. Chapter 11. Baltimore: Paul H. Brookes Publishing, 127-139.	■ Bayley Mental Indices ■ Stanford-Binet ■ WPPSI ■ WISC-R ■ Peabody Individual Achievement Test (PIAT)	■ No indication limited English speaking children were involved. ■ PIAT given in kindergarten and following year.

	Name of Program	Reference	Outcome Measures	Comments
B.15	Family Development Research Program/ Syracuse University (1970s)	Lally, J.R., Mangione, P.L., & Honig, A.S. (1987, September). The Syracuse University Family Development Research Program: Long-range impact of an early intervention with low-income children and their families. San Francisco: Far West Laboratory.	Cognitive functioning: ■ Stanford Binet Social-Emotional: ■ Observations using the Social-Emotional Observer Ratings of Children	■ No indication limited English speaking children were involved.
B.16	High/Scope Preschool Curriculum Comparison Study (late 1960s, early 1970s, and later follow-up)	Schweinhart, L. et al. (1986). Preschool curriculum comparison study. Ypsilanti, MI: High/Scope Educational Research Foundation.	Domain not specified: ■ Stanford Binet IQ ■ California Achievement Test	■ Children tested at ages 3,4,5 and 6 (kindergarten); CAT given in first grade. ■ No indication that limited English proficient children included in sample.

C. OTHER DATA COLLECTION EFFORTS

Name of Program	Reference	Outcome Measures	Comments
C.1 National Longitudinal Surveys of Youth Child Assessments (1986)	NLS Handbook (1988). Columbus: Center for Human Resource Research, the Ohio State University, 99-102. Olsen, R.J. (1989, Spring). New databases in human resources. The national longitudinal surveys of labor market experience merged child-mother data, 24(2), 336-339.	Nature and quality of child's developmental environment: ■ Maternal self-reports and interviewer observations using Home Observation for Measurement of the Environment, Short Form (HOME-SF) Oral verbal knowledge/vocabulary: ■ Body Parts Receptive vocabulary of standard American English: ■ Peabody Picture Vocabulary Test-Revised (PPVT-R) Short-term memory: ■ Memory for Location ■ McCarthy Scale of Children's Abilities; Verbal Memory Subscale ■ WISC-R; Digit Span Subscale (for children aged 7 and older) Mathematics: ■ Peabody Individual Ach. Test (PIAT); Math Subscale (for children aged 5 and older) Oral reading: ■ PIAT; Reading Recognition Subscale (for children aged 5 and older) Reading comprehension: ■ PIAT; Reading Comprehension Subscale (for children aged 5 and older)	■ Assessments completed with 4,971 children aged below 11 and all maternal ages below 28. Mothers considered to be a nationally representative cross-section of women aged 21 to 28 on 1/1/86 or 14 to 21 on 1/1/79; blacks, Hispanics, economically disadvantaged whites, and military personnel were over-sampled. Limited English speaking children were either excluded from testing or assessed in English.

Name of Program	Reference	Outcome Measures	Comments
C.1 National Longitudinal Surveys of Youth Child Assessments (1986) (Continued)		Behavioral style of children: ■ Maternal report using Temperament Scales ■ Maternal report using Behavior Problems Index ■ Interviewer ratings of child behavior in testing situation Self-worth: ■ Child's self-report using Perceived Competence Scale for Children/Self-Perception Profile (children aged 8 and older) Developmental milestones on motor, cognitive, communication, and social development: ■ Motor and Social Development Scale	

APPENDIX B

STATUS OF COGNITIVE AND SOCIAL-EMOTIONAL MEASURES ON CRITERIA FOR INITIAL SCREENING

TABLE B.1: PRELIMINARY SCREENING OF MEASURES FOR THE COGNITIVE DOMAIN

	Relevance to Domain	Relevance to Age Span	Administration Time	Training of Examiners	Availability in Other Lang.
Adaptive Language Inventory (Feagans & Farron, 1979)	?	?	+ (untimed for adult to complete)	+ (typically used by caregiver)	N/A
Battelle Dev. Inventory-Screening Test	+ (also covers social-emotional)	+ (infancy-8 yrs.)	+/- (20-30 min.)	+ (paraprof.)	N/A
Brigance Screen (both Preschool & K-1 versions)	+	+ (ages 3-4 and grades K & 1)	+ (15 min.)	+/- (paraprof., requires judgements)	+ (Spanish)
California Achievement	+	- (grades K-12)	- (2+ hrs.)	+ (paraprof.)	?
Chicago EARLY Assessment	+	+ (ages 3 to 6)	+ (15-20 minutes)	+ (paraprof.)	+ (Spanish)
Children's Embedded Figures Test	- (too narrow)	- (ages 5-12)	+/- (15-30 min.)	+ (paraprof.)	?
Denver Developmental Screening Test	+ (also covers social-emotional)	+ (ages 0 to 6)	+/- (20 min.)	- (non-standard admin.)	?
DIAL-R	+	- (ages 2 to 5-11)	+/- (20-30 min.)	+ (paraprof.)	?
Early Screening Inventory (ESI)	+	+ (ages 4-6)	+ (15-20 min.)	+ (paraprof. with background in child dev.)	?
Head Start Meas. Battery	+	+	- (120 min.)	+ (paraprof.)	+ (Spanish)
Illinois Test of Psy. Ability	+	+ (ages 2-10)	- (60 min.)	+/- (child dev. training)	+ (Spanish version available from author)
Iowa Test of Basic Skills (ITBS)	+	- (grades K-9)	- (150-235 min.)	+	?
Kaufman Assessment Battery (K-ABC)	+	+ (ages 2.5-12.5)	- (60 min.)	+ (paraprof.)	+ (Spanish)
Kindergarten Performance Profile (Brookline)	+ (also covers social-emotional)	- (grades K-2)	+ (untimed for adult to complete; 10 min.)	+ (typically used by teacher)	N/A
Language Assessment Scales (LAS)	- (too narrow; meas. oral language)	+/- (grades K-12)	+ (10-20 min. dependent on form)	+ (paraprof.)	+ (Spanish)
McCarthy Scale of Children's Abilities (Verbal Memory Scale)	+	+ (ages 2.5-8.5)	+/- (60 min.; selected subtests less)	+/- (paraprof. with child dev. training)	?

BEST COPY AVAILABLE

TABLE B.1: PRELIMINARY SCREENING OF MEASURES FOR THE COGNITIVE DOMAIN (CONTINUED)

	Relevance to Domain	Relevance to Age Span	Administration Time	Training of Examiners	Availability in Other Lang.
Metropolitan Readiness	+	+ (grades pre-k to 1st)	- (95 min. for level 1; 80 min. for level 2)	+/- (typically the classroom teacher)	?
PDC Child Interview (Adapted from Purdue Social Attitude Scale)	- (too narrow; meas. interest in reading and attitude toward school)	+/- (grades 1-3)	+ (< 20 min.)	+ (paraprof.)	+ (Spanish)
PDC Parent Interview	- (too narrow; meas. child's attitude toward school)	+/- (grades 1-3)	+ (< 20 min.)	+ (paraprof.)	+ (Spanish)
Peabody Picture Voc. Test (Revised)	- (narrow)	+ (ages 2.5-40)	+/- (15-20 min.)	+ (paraprof. with training in child dev.)	+ (Spanish)
Peabody Indiv. Ach. Test (Reading & Math)	+	- (grades K-12)	+/- (30-50 min. for all 5 subtests; reading and math only would take less time)	+ (paraprof.)	?
Preschool Inventory, Version R with 32 items (PSI)	+	+ (ages 3-5)	+ (15 min.)	+ (paraprof.)	+ (Spanish)
Preschool Language Assessment Instrument	- (narrow)	+ (ages 0-7)	+ (20 min.)	?	?
*Raven's Coloured Progressive Matrices	- (too narrow; meas. non-verbal problem-solving)	?	?	?	?
*Sigel Object Categorizing Test					
Walker Readiness Test	- (may be too narrow; meas. verbal and school readiness)	+ (ages 4-6)	+ (untimed; 8-10 min.)	+ (paraprof.)	?
Wide Range Achievement Test (WRAT)	+	- (ages 5-adult)	- (30 min.)	- (educ. or psy. prof.)	?
WPPSI (Block Design Subtest)	- (narrow)	+ (ages 4 to 6.5)	+/- (75 min. for 11 subtests, indiv. subtests would be less)	+/- (Block Design Subtest could be admin. by a trained paraprof.)	?
Zimmerman Preschool Lang. Scale (PLS)	- (narrow; meas. school readiness of integrated auditory and visual perceptual modes)	+ (ages 2-9)	+ (15 minutes)	- (difficult to score)	?

*Additional screening information will be added as it becomes available.

TABLE B.2: PRELIMINARY SCREENING OF MEASURES FOR THE SOCIAL-EMOTIONAL DOMAIN

	Relevance to Domain	Relevance to Age Span	Administration Time	Training of Examiners	Availability in Other Lang.
Behavior Problems Index (Based on the Child Behavior Checklist and Revised Child Behavior Profile by Achenbach)	- (too narrow; meas. severe beh. problems, aggression)	+	+ (untimed for adult to complete; <10 min.)	+ (typically used by caregiver)	N/A
Brown Self-Concept Test	- (may be too narrow; meas. feelings about self and school)	+	+ (< 20 min.)	?	?
Calif. Preschool Competency Scale	+	+/- (ages 2.6 - 5.6)	+ (untimed for adult to complete)	+ (typically used by teacher)	N/A
Child Behavior Rating Scale -- Used in Home-based Study (Based on Battelle Dev. Inventory)	+	+ (grades Pre-K to 5)	+ (untimed for adult to complete; 30-40 min.)	+ (typically used by teacher)	N/A
Child Behavior Rating Scale -- Used in Giant Step (Based on Bronson Exec. Skill Profile)	+	+ (ages 3-5)	+ (untimed for adult to complete; 40-40 min.)	+ (typically used by teacher)	N/A
*Classroom Behavior Inventory					
Hahneman Elem. School Behav.	+	- (for use with elementary aged children)	+ (untimed for adult to complete; 15 min.)	+ (typically used by teacher)	N/A
Head Start Meas. Battery (Social Scale)	- (too narrow)	+	?	+ (paraprof.)	?
High/Scope Pupil Obs. Checklist (Adapted from Pupil Obs. Checklist)	+/- (Narrow; meas. of task orient., sociability)	+ (grade Pre-K to 3)	+ (untimed, for adult to complete; 20 min.)	+ (typically used by teacher or examiner)	N/A
*Intellectual Ach. Responsibility Scale (IARS)	- (too narrow; meas. locus of control)	-			
Kohn Social Competence Scale	- (narrow)	+ (3-6 years)	+ (untimed for adult to complete; 15 min.)	+ (typically used by teacher)	N/A

*Additional screening information will be added as it becomes available.

TABLE B.2: PRELIMINARY SCREENING OF MEASURES FOR THE SOCIAL-EMOTIONAL DOMAIN (CONTINUED)

	Relevance to Domain	Relevance to Age Span	Administration Time	Training of Examiners	Availability in Other Lang.
PDC Child Rating Scale	+/- (may be too narrow; meas. learning orient., independence, and social adjustment)	- (grades 1-3)	+ (untimed for adult to complete; 20 min.)	+ (typically used by teacher)	N/A
*Perceived Competence Scale for Children/Self Perception Profile					
Preschool Behavior Questionnaire (PBQ) (Behar & Stringfield, 1974)	-/+ (narrow, meas. maladjustment)	+ (ages 3-6)	+ (untimed for adult to complete; 10 min.)	+ (typically used by teacher)	?
Preschool Interpersonal Problem Solving Test (PIPS) (Shure & Spivak, 1974)	- (too narrow; meas. interpersonal problem-solving skills)	- (ages 4-5)	+ (untimed; < 20 minutes)	- (examiners must judge child's responses)	?
Pupil Observation Checklist (POCL)	- (narrow; meas. test-taking beh.)	+ (ages 3-6)	+ (5 min. for adult to complete)	+ (paraprof.)	N/A
*Purdue Social Attitude Scale					
Schaefer Behavior Inventory	+/- (may be too narrow; meas. task orientation, extro-version/intro-version, hostility/ tolerance)	+	+ (untimed; < 20 min. for adult)	+ (typically used by parent or teacher)	N/A
*Social Obs. Code (Poteat & Saudargus)					
Stephens/Delys Rein. Contingency Interview	+	+ (ages 4-10)	+ (10-25 min.)	+ (trained paraprof.)	?
Values Inventory for Children	+/- (sociability, me first, moral values)	- (grades 1-7)	?	+/- (teacher)	?
*Waters & Deane Attachment Q-Set					

*Additional screening information will be added as it becomes available.

TABLE B.3: PRELIMINARY SCREENING OF CHILD OBSERVATION INSTRUMENTS

	Relevance to Domain	Relevance to Age Span	Administration Time	Training of Observers	Availability in Other Lang.
Beller Rating Scales	- (may be too narrow; meas. dependency, aggression)	+	+/- (untimed for observer to complete; six, 15 min. obs.)	+ (trained observer)	N/A
Bronson Social and Task Skill Profile	+ (meas. use of time, mastery beh., and social beh.)	+ (ages 3-8)	+/- (untimed for observer; six, 10 min. obs.)	+/- (trained observer but complex)	N/A
Howes Peer Play Scale	+/- (meas. social interaction and friendships with peers)	pre-K	+/- (untimed for observer; four, 5 min. obs.)	+ (trained observer)	N/A
Personal-Social Behavior Rating Scales (Emmerich, 1971)	+/- (maybe too narrow; meas. aspects of personality)	+	+ (20-30 min.)	+ (trained observer)	N/A

APPENDIX C

PROFILES OF INSTRUMENTS MEETING PRELIMINARY CRITERIA

Battelle Developmental Inventory Screening Test (BATTELLE-S)

Publisher/Date: DLM Teaching Resources (1984)

Description: The BATTELLE-S contains 96 items selected to represent the contents of the full scale, the Battelle Developmental Inventory (BDI). The BATTELLE-S is designed to assess skills in five domains: personal/social, adaptive, motor (gross motor and fine motor), communication (receptive and expressive), and cognition. The BATTELLE-S uses three procedures to collect test data: (1) structured test administrations; (2) interviews with parents and/or teachers; and (3) observations.

Technical Characteristics:

Validity:

1. Predictive validity of the BATTELLE-S is based on its relationship to scores on the full scale BDI. A study of 164 children who were tested with both the BATTELLE-S and the BDI yielded correlation coefficients of .92 and above for the 10 scores on the BATTELLE-S and comparative BDI components.

2. Experts agreed the BDI was content valid.

3. Construct validity established for specific conditions (clinical versus nonclinical) and constructs (factors and age); all correlations above .81.

Reliability:

1. Internal consistency not calculated.

2. Coefficient for test/retest reliability was .99 (total sample); and .99 for inter-rater reliability.

Norms:

1. The BDI was nationally standardized in 1982/83 with a stratified sample of 800 children. Characteristics of the sample reported for age, sex, race and geographic region. The minority group was 8.9 percent Black, 6.4 percent Spanish origin, and 0.7 percent other. Sample was approximately 75 percent urban and 25 percent rural with an emphasis on middle SES. Data for the BATTELLE-S were collected as part of the norming process for the full Inventory.

Practical Considerations:

Compatibility:

The BATTELLE-S is listed in a bibliography of tests recommended for use in Chapter 1 early childhood programs.

Administration:

1. The test takes 20-30 minutes for children between ages 3-5 years, and about 10-15 minutes for children under 3 years or over 5 years.

2. The examiner administers many of the structured items to the child while both are seated at a child-size table. The examiner uses basal and ceiling levels to determine the items in each domain/subdomain to be administered to the child. The examiner orally presents the item stimulus and the child's responses may be oral, pointing, or motor, as appropriate.

3. Examiner training, including test familiarization and practice in administration, is required. For each test item, the examiner is provided with detailed instructions for specific behavior to be assessed. required materials, assessment procedures, and scoring criteria. Test can be administered by a teacher or trained paraprofessional.

Scoring:

1. The child's performance on each item may be scored 0 (incorrect or no response); 1 (attempted but did not fulfill all criteria); or 2 (met all criteria. Raw scores are calculated for the five domains, the four subdomains, and the total test. Cut-off scores and age equivalent scores are also provided for the total test domains and subdomains.

Language Fairness:

Publisher indicated that the BATTELLE-S is not available in languages other than English.

Effectiveness:

Previous use:

Pfannenstiel, J.C. & Seltzer, D.A. (1985).
Evaluation Report: New Parents as Teachers (NPAT).
Overland Park, KS: Research and Training Associates.

Authors of NPAT evaluation report that parents used the "personal-social" domain of the BDI to assess their child's social development. The results of a factor analysis revealed significant differences between treatment/control groups on four of six scales.

Brigance Preschool and K-1 Screen

Publisher/Date: Curriculum Associates (1985)

Description: The Brigance Preschool and K-1 Screen are two criterion referenced developmental screening tests. The Preschool Screen evaluates basic developmental and readiness skills of three- and four-year-old children. The K-1 Screen assesses the basic skills necessary for success in grades K-1. Eleven basic assessments are included in each version: general knowledge and comprehension, speech and language, gross motor skills, fine motor skills, and math.

Children complete multiple oral-response and task-performance items on both versions. The K-1 Screen also includes paper-pencil items and direct-observation assessments. Personal information, assessment results, scoring, testing observations, comparative summary of the screening, and recommendations are all recorded on the pupil data sheet. Test items on the K-1 Screen are cross-referenced to the Brigance Inventory of Basic Skills.

Technical Characteristics:

Validity:

1. Content validity is based on a literature review and field testing conducted in 12 states.

2. Technical data on criterion related or construct validity not available.

Reliability:

1. Technical data on reliability not available.

Norms:

1. Norms are not available.

Practice Considerations:

Compatibility:

The Brigance Screen is listed in a bibliography of tests recommended for use in Chapter 1 Early Childhood Programs. The instruments are more appropriate for use in screening children for placement and/or referral for further assessment.

Administration:

1. Administration time for each version generally ranges from 12 to 15 minutes.

2. Both versions may be administered individually by one examiner or by a team of examiners at multiple stations. For each assessment, the examiners' detailed directions for administration and scoring are presented in a standard format. The child's required responses may be oral, pointing, motor, or written. Alternative methods include the use of teacher's ratings, parent's ratings, or data from school records.

3. The examiner may be a teacher or trained paraprofessional.

Scoring:

The correct response to each item earns from 2 - 5 points, depending upon the assessment. The examiner calculates the score for each assessment and then sums them to obtain the total score. The total possible score is 100. The scores for all children are ranked and assigned to three groups: average, lower than average, and above average. The manual presents an example for establishing a cut-off score, using the lower 20 percent of the group.

Language Fairness:

Publisher indicated that directions and questions the examiner reads to the child are available in Spanish.

Effectiveness:

Jarvis, C.H., Molnar, J.M. & Collins, C. (1985). Urban education: Can all-day kindergarten make a difference? Paper presented at the annual meeting of the American Psychological Association, Los Angeles, CA.

Fall and spring scores on the Brigance K & 1 Screen were one of two measures used with English and non-English speaking children (primarily Spanish) to determine whether a longer school day would have measurable effects on student growth. Brigance scores yielded significant main effects for kindergarten (full vs. half-day) and home language (English vs. language-minority) groups. A significant interaction between the length of the kindergarten program and home language was found.

The author reported several disadvantages in using the Brigance: (1) it is a criterion-referenced instrument with no national or local norms (a control group would overcome this problem); (2) the absence of published technical data; and (3) the K-1 Screen has a low ceiling.

Locally established reliability data regarding the Brigance was included in the evaluation findings. Test-retest stability of items in each skill category ranged from .43 to 1.00 with a total test-retest reliability of .91. The inter-item consistency ranged from .38 to .94 on nine skill categories. When all 72 items were combined, internal consistency was .91.

Bronson Social and Task Skill Profile

Developer/Date: Martha B. Bronson (1985)

Description: The Bronson is an observational measure that uses structured observation categories, a modified time sampling technique, and trained observers to assess the behavior of individual children between the ages of three and eight years.

The Bronson records observations of an individual child in 10 minute segments. During the 10 minutes four areas of the child's activity are described and recorded in terms of frequency and duration:

Use of time, including several categories of social and non-social activities;

Mastery behaviors, including categories of behaviors that are positively or negatively related to competence in mastery tasks;

Social behaviors, including categories for behaviors that are positively or negatively related to social competence.

An **activities section** is used to write a brief narrative record of the ongoing activities of the child; while specific actions, interactions, and the object of interactions are noted next to the checks and letter codes in specific scoring categories.

Theoretical Base: The Profile provides a way of evaluating children's social behaviors, mastery behaviors, and their use of time, all within the natural setting. The underlying hypothesis is that the concept of "executive" ability or skills can be applied to all three areas of performance. The term is used in the information processing sense of "executive routines" or "programs" (plans) which organize and guide behavior. Executive skill implies skill in recognizing the relevant cues, parameters, or rules, of a situation; skill in predicting and planning possible sequences of events and outcomes of a situation; and skill in organizing and controlling both the self and the social or material "other" in a situation in order to effectively reach chosen goals.

Technical Characteristics: Validity:

1. Construct validity established through intercorrelations among 11 competence variables measured for fall and spring of kindergarten (N = 358), and the spring of the second grade year (N = 408). Intercorrelations were in the expected directions.

2. Concurrent validity. Kindergarten children rated by their teachers having high or low general competency were also observed using 11 selected competency variables from the Profile. Means of the "low" rated children were more than a standard deviation below the "high" rated group on the mastery variables. Differences were smaller in the social and use of time variables, but consistently favored the "high" group except in a "rate of social acts" variable.

3. Predictive validity. Between 46 and 64 percent of all children having observed problems in kindergarten also had observed problems in the second grade. Spring problems were a slightly better predictor than fall problems, and children with problems in both fall and spring were most likely to have observed problems in the second grade. The predictive validity of the observations compared with other established predictors (low cognitive test scores, low mother's education, and male sex) using simple correlations ranged from .29 to .41.

Reliability:

1. In pilot studies, stability was highest in the mastery variables (.60 with a range of .54 to .69), lower in the social variables (.27 with a range of .10 to .47), and lowest in the use of time variables (.19 with a range of .04 to .26).

Norms:

Author notes that means and standard deviations of the various pilot groups observed cannot be considered normative for children with different demographic characteristics.

Practical Considerations:

Compatibility:

For this observation system to be useful, there must be sufficient free play or free choice time for observers to be able to complete their observations. Highly structured and/or teacher-directed classrooms may limit the opportunities for data collection with this instrument.

Administration:

1. The length of the observations can vary, but 10-minute observations have typically been used in classroom settings for each of the six observations. A modified time sampling method is used in which three of the observations are started at the beginning of a social interaction and three are started at the beginning of a mastery task. Observations should be scheduled when children have some free choice about which activities to engage in, and all observations should be done in the classroom setting itself (not in the lunchroom, outside, etc.).

2. Preschool and kindergarten mastery observations should be done during tasks with a recognizable goal and recognizable steps to be used in reaching the goal (puzzles, matching and sorting, etc.) and when the child is working independently.

3. Each of the six observations on a single child should be done on a different day over a period of no less than two weeks and no more than six weeks. Though no more than one observation per day per child is considered optimum, observers can do one mastery and one social observation per day per child when pressed for time.

4. Observers must have skill in unobtrusive observing and be able to keep track of the ten-minute observation times in intervals of 15 seconds. Observers must be trained to note specific actions, interactions, and the objects of interactions, as well as to write a detailed narrative description at the end of each ten-minute observation. The Profile may be adapted to require the observation of fewer types of actions/interactions.

Scoring:

1. For scoring, data from the separate observations are pooled for a particular child within each of the categories. A numerical rate or percent score is obtained for each of the behavior activities observed.

Language Fairness:

Not applicable.

Effectiveness:

The Profile is being used by Abt Associates in an evaluation of Project Giant Step in New York City.

The Profile was used as part of the evaluation of the Brookline Early Education Project (BEEP), and many of the technical characteristics of this instrument were established using data from the BEEP evaluation.

Child Behavior Rating Scale (CBRS-1)
(Based on the Battelle Developmental Inventory)

Publisher/Date: Created by RMC Research Corporation (1986)

Description: This 35-item instrument is completed by the child's teacher or home visitor. Based on the Personal-Social and Adaptive Scales of the Battelle Developmental Inventory, this rating scale addresses adult interaction, expression of affect, peer interaction, coping, social role, self-concept, and task mastery for children between the ages of 3 and 5.

Technical Characteristics:

Validity:

Measures of content, criterion, or construct validity not available.

Reliability:

Internal consistency (Cronbach: Alpha) of .92 found at pretest in Home-based study.

Norms:

Not available.

Practical Considerations:

Compatibility:

Child behaviors that form the basis of this rating scale are congruent with more child-directed or experiential classroom settings used by Chapter 1 preschool programs.

Administration:

1. This version of the CBRS is an untimed paper-pencil rating scale that takes a teacher approximately 10-20 minutes to complete per child.

2. The scale is easy to complete and no special training is needed; however, the rater needs to be well acquainted with the child.

Scoring:

1. The rater must circle a number on a four-point scale -- "not at all like," "somewhat like," "much like," and "very much like" -- to indicate how similar the child is to the behavior described in a particular item. The rater can also indicate that there has been no opportunity for the child to demonstrate the particular behavior or the behavior has not been observed. A child's score is the mean rating across the 35 items.

Language Fairness:

Not applicable; teacher completes the rating scale.

Effectiveness:

Meleen, P., Love, J., Nauta, M. (1988). Study of the home-based option in Head Start, Final Report, Vol. 1: Technical Report. Hampton, NH: RMC Research Corporation.

Pre- and post-testing on the CBRS yielded no major differences in the effectiveness of a particular service delivery mode. In all groups, children did show gains in areas of social development from pre- to post-testing. Failure to find significant program impacts may have been due to ceiling effects.

Child Behavior Rating Scale (CBRS-2)
(Based on the Bronson Executive Skill Profile and the RMC Research CBRS)

Publisher/Date: Adaption by Martha Bronson and John Love (Abt Associates, 1986)

Description: The 34 items on this rating scale were based on coding categories -- from the RMC CBRS used in the Home-based evaluation and Bronson Executive Skill Profile Observation System. The adapted CBRS evaluates a child's social behavior with peers, with adults, and task behavior.

Technical Characteristics:

Validity:

1. In the Giant Step evaluation rating scale, items on task orientation/strategies were more strongly correlated with Preschool Inventory scores than were the adult and peer interaction items.

Reliability:

1. Fall-spring correlation of .67 for ratings of 364 Giant Step children by their teachers.

2. Internal consistency (Cronbach's Alpha) reported as .96 overall. Subscales were individually reported as .90 for interaction with peers, .70 for adult interactions, and .96 for task behavior.

Norms:

Not available.

Practical Considerations:

Compatibility:

Aspects of personal-social behavior evaluated by this version of the CBRS are congruent with Chapter 1 preschool classroom settings.

Administration:

1. This version of the CBRS takes 10-15 minutes for an adult to complete regarding one child.

2. The scale is a 34-item paper-pencil rating scale typically completed by the child's teacher.

3. The scale is easy to complete and no special training is required. The rater must be familiar with the child and how he/she interacts.

Scoring:

1. The rater circles a number on a five-point scale to indicate how frequently the behavior occurs -- "The child never/rarely/sometimes/ frequently or usually/always) exhibits the behavior described by the item." In addition, the teacher is asked to estimate the percentage of time the child spends on four types of activities when the child can choose what to do (social play, working with materials, solitary fantasy play, or monitoring/uninvolved).

2. Total score is the mean rating of all items rated.

3. Three subscores available: child interactions, adult interactions, task behavior.

Language Fairness:

Not Applicable since teacher completes the rating scale.

Effectiveness:

Abt Associates (1988, October). Evaluation of Project Giant Step: Technical Progress Report Number 4. Cambridge, MA: Abt Associates.

Across all Giant Step centers, ratings improved from fall to spring.

California Preschool Social Competency Scale (CPSCS)

Publisher/Date: Consulting Psychologists Press (1969)

Description: The California Preschool Competency Scale assesses the social competency of preschool children ages 2.6 - 5.6. It is typically used by preschool teachers for diagnosis, placement, or measurement of the development of young children. This is a 30-item paper-pencil rating scale used to rate interpersonal behavior and the degree to which children assume social responsibility. Specific behaviors that are rated include using names of others, following verbal instruction, sharing, and accepting limits.

Technical Characteristics:

Validity:

Test reported to have face validity; no information on predictive validity and there is no recognized standard of social competence with which it could be compared.

Reliability:

1. Inter-rater reliability ranges from .75 to .86, with split-half reliabilities above .90.

2. High internal consistency (odd-even reliability of .96); stability not reported.

Norms:

A manual provides separate percentile norms for boys and girls and for high and low socioeconomic groups. One reviewer (Robert Calfee from Stanford) has noted that boys and preschoolers from a low-income family match relatively poorly the expectations of the preschool teachers represented in the norming of the scale. Calfee reasoned that such children may be quite well adjusted to the social demands of their environment and may match adequately the standards of preschool teachers of a persuasion different from those who generated the scale. Another reviewer has questioned the adequacy of the norming data because there are 16 sets of norms, and each one is based upon only 50 children. Mediax Associates (1980) reported the CPSCS has been normed using primarily middle class children. Some test items may be culturally-biased.

Practical Considerations:

Compatibility:

The social-emotional areas measured by the CPSCS are relevant to the classroom environment of Chapter 1 preschool programs.

Administration:

1. The CPSCS is untimed.

2. The test is a 30-item paper-pencil rating scale completed by teachers.

3. The scale is easy to complete and no special training is needed; however, the rater needs to be well acquainted with the child.

Scoring:

1. The test provides numerical evaluations of the social competency of children. Every item is rated on a 4-point scale arranged in order of increasing competency. The total social competency raw score is the sum of the ratings for the 30 items.

Language Fairness:

Not applicable since teacher completes the rating scale.

Effectiveness:

Lee, V.E. et al., (1989). Are Head Start effects sustained? Unpublished manuscript.

Author reported measures used in the original Head Start Longitudinal Study (1972 data). The present re-analysis of the data found significant effects in social competence favoring girls. A Head Start effect (compared to no preschool) was found favoring males.

Howes Peer Play Scale

Developer/Date: Carollee Howes (1980; 1987)

Description: The Peer Play Scale is a classroom observation instrument that is designed to measure peer interactions and friendships of children ages one to six. This instrument was developed by Carollee Howes and has been used with preschool children in child care centers. The 1987 revised version of the Peer Play Scale measures solitary behavior and proximity of peer partners; parallel activity with or without no awareness of peers; simple social play, and complementary and reciprocal play; cooperative and complex social pretend play; and attempts to play games with rules (e.g., football, checkers).

Theoretical Base: Howes defines social competence as behavior that reflects successful social functioning with peers. This definition of competence includes two independent yet related aspects: (1) social interaction skills and (2) friendships. Social interaction skills include ease of entry into play groups, play with peers, affective expressions, and other behaviors that lead to peer acceptance and popularity. Friendships are defined as stable, dyadic relationships marked by reciprocity and shared positive effects.

Howes has developed this observational system based on three assumptions: (1) the specified sequence of behavioral constructs remains constant across children with variations in their experiences with peers and social relationships with adults; (2) variations in the behavioral construct used to represent social competence within each developmental period correspond to variations in the social competence of the children; and (3) individual differences in social competence remain stable across developmental periods.

Technical Characteristics: Validity:

1. Pearson product-moment correlations between observed behaviors and teacher ratings of sociability with peers were moderate to high. These correlations decreased in strength with the children's age. Teacher ratings of peer relationships correlated moderately with observed attempts to initiate play in younger age groups, but not in the four to six-year-olds.

Reliability:

1. Stability of the observed measures ranged from .70 to .91 across observational sessions over a four-week period.

2. Indices of intercoder reliability were computed using kappa coefficients. All indices of intercoder reliability were above .89.

Norms:

No information provided.

Practical Considerations:

Compatibility:

This instrument is used by observers during free play periods. Highly structured and/or teacher-directed classrooms will limit opportunities for data collection.

Administration:

1. Each child is observed four times in random order during free play periods. An observation begins when a child begins to interact with a peer and continues for five minutes regardless if the child contiues to interact with peers. Interaction is defined as social behaviors directed to or from the target child and a peer partner, or involvement in a mutual game.

2. Instructions for the observer and data collection forms are not provided. The author would need to be contacted to determine if draft copies are available.

Scoring:

1. Actual scoring procedures do not accompany the description of the instrument or a monograph that reviews technical characteristics. The proportion of time spent in each type of play situation is computed as individual scores.

Effectiveness:

The Howes Peer Play Scale was used as part of the National Child Care Staffing Study completed by M. Whitebook, C. Howes, and D. Phillips in 1989. The scale performed as expected.

Howes, C. (1987). Social competence with peers in young children: Developmental sequences. Developmental Review, 7, 252-272.

Howes, C. (1980). Peer play scale as an index of complexity of peer interaction. Developmental Psychology, 16, 371-372.

McCarthy Scales of Children's Abilities (MSCA)

Publisher/Date: The Psychological Corporation (1972).

Description: The purpose of the McCarthy is to predict a child's ability to cope with school work in the early grades. Six scales (18 component tasks) measure the following abilities: right/left orientation, verbal memory, draw-a-person, numerical memory, conceptual grouping, and leg coordination. A shorter version of the MSCA is called the McCarthy Screening Test. It also contains tasks in the same six areas.

Technical Characteristics:

Validity:

1. The six MSCA scale correlations with other ability tests range from .62 to .71 (with the Stanford Binet), and .27 to .61 (with the WPPSI-IQ).

2. Predictive validity coefficients range from -0.7 to .57 for individual scales from the MSCA and from .34 to .54 on the general cognitive scale. The Metropolitan Achievement Test was used to establish predictive validity.

Reliability:

1. Stability coefficients for the six MSCA scales for all age groupings range from .69 to .91. The general cognitive scale ranges from .89 to .91 for the three age groupings.

2. Intercorrelation of the six MSCA scales range from .37 to .95. The general cognitive scale ranges from .80 to .95. Higher intercorrelations between scales is attributed to high overlapping content.

Norms:

The standardization of the MSCA was based on a nationwide sample that was stratified on several major variables. Stratification variables used include age, sex, color, geographic region, father's occupation, and urban-rural residence. Bilingual children were eligible for testing only if they could speak and understand English. As part of the standardization process, a weighted raw score for each scale was determined for each child in the standardization sample. These raw scores were then converted to scaled scores and resulting norms were then arranged in sequence by age group.

Practical Considerations:

Compatibility:

The MSCA scales are very compatible with approaches being used in Chapter 1 preschool programs.

Administration:

1. The complete MSCA takes approximately 60 minutes (less than 10 minutes per scale); a shorter version, the McCarthy Screening Test takes approximately 20 minutes. Both tests are untimed.

2. Except for the verbal memory scale, of which only Part 1 is given, each of the tests is administered in its entirety. The child's required responses may be oral or motor, as appropriate.

3. The MSCA requires at least a paraprofessional with background in child assessment and child development. The publisher stresses that the examiner should be clinically familiar with the MSCA battery. Instructions for administering the battery are quite detailed, but still require judgement in scoring the accuracy of a response.

Language Fairness:

The publisher indicates the MSCA is not available in languages other than English.

Effectiveness:

Bond, J.T. et al., (1982). Project developmental continuity evaluation final report. Ypsilanti, MI: High/Scope Educational Research Foundation.

Only test items from the verbal and perceptual-performance scales were used. Spanish dominant children were excluded from the analysis. Test items did not yield significant positive effects across sites.

NLS Handbook (1988). Columbus: Center for Human Resource Research, the Ohio State University.

Only test items from the verbal memory subscale were used. Non-English speaking children were included in the sample.

Peabody Picture Vocabulary Test-Revised (PPVT-R)

Publisher/Date: American Guidance Service (1981)

Description: The PPVT-R consists of two forms. The test allows a verbal or nonverbal response by the child and is untimed. A child is asked to indicate which of four pictures presented on a carousel-mounted plate corresponds to a stimulus word read aloud by an examiner. The test measures receptive vocabulary in English.

Technical Characteristics:

Validity:

1. A comparison of scores from the PPVT and other child IQ measures revealed correlations of .82 and .96. PPVT IQ scores correlated with WISC-R from .30 to .84. The publishers have concluded that the PPVT-R is not a comprehensive measure of IQ, but that it does help predict school success.

Reliability:

1. Numerous studies have demonstrated the reliability of the PPVT-R.

Norms:

The PPVT-R norms are based on a nationwide sample representative of the U.S. population according to the 1970 census. Minorities were included in the standardization sample and sex or ethnic stereotyping was eliminated.

The Spanish version of the PPVT-R, called the Test de Vocabulario en Imagines Peabody (TVIP), has the same structure and standard score system. Separate standardizations were conducted with Spanish-speaking children in Mexico and Puerto Rico. Both combined and separate norms are available to interpret results.

Practical Considerations:

Compatibilty:

The PPVT-R and TVIP are compatible with the language focus taken in many Chapter 1 Early Childhood Programs, but do not address other cognitive areas relevant to child development and school success.

Administration:

1. Both versions of the test take 15-20 minutes.

2. Administration procedures require the child to respond only to the items between the basal and ceiling. To administer the scale, the examiner shows a plate containing four pictures arranged in a multiple choice format and says the corresponding stimulus word. The child points to the picture which best illustrates the meaning of the stimulus word.

3. The examiner may be a trained paraprofessional.

Scoring:

A score is obtained by subtracting errors from the total ceiling score and may be converted to percentile rank, age equivalent score, or a standard score.

Language Fairness:

The TVIP permits the assessment of Spanish-speaking children in their first language.

Effectiveness: All or part of the PPVT-R has been used in the following studies:

- The Child Care Staffing Study:
- Project Developmental Continuity Evaluation (High/Scope).
- Home Start Follow-up Study (Abt Associates);
- National Day Care Study (Abt Associates):
- Head Start Planned Variation Study (used PPVT);
- The At-Risk Preschool Program (Chicago Public Schools);
- The Pre-K Program (Austin Indep. School District):
- Daycare Programs for Disadvantaged (Bermuda);
- Pre-K Program (New York State)
- National Longitudinal Survey of Youth/Child Assessments (Center for Human Resource Research, the Ohio State University).

The PPVT-R has performed well consistently. It has, however, usually been used as part of a larger battery of tests measuring cognitive ability.

Preschool Inventory - Revised (PSI)

Developer/Date: Educational Testing Service (1976)

Description: The PSI was developed originally by Bettye Caldwell to provide Head Start with a practical measure of preschool achievement. It was designed to measure educational achievement (e.g., child's knowledge of basic concepts such as first/last, under/behind, colors, shapes, and knowledge of body parts). The PSI uses a structured testing setting in which the examiner orally presents the test items and the child's responses may be oral, pointing, or motor, as appropriate.

Technical Characteristics: Validity:

1. PSI test scores reported as correlating .45-.56 with each of five age groups from the standardization sample. Correlations between PSI test scores and Stanford Binet Intelligence Test scores for 1476 children in the standardization sample ranged (by age group) from .39 to .65, with .44 being the correlation for the entire sample.

2. PSI test scores reported as correlating .42 with ratings on the Coleman Index and .50 with scores on the Home Information Scales. These two measures of SES reported as correlating at .51 with each other (data taken from a study in North Carolina that included 317 children in eight kindergarten centers).

Reliability (based on earlier versions of PSI):

1. Split-half reliability (internal consistency), corrected by the Spearman-Brown formula reported as .95 on an earlier version (64-item) of the PSI.

Norms:

The PSI was initially standardized with 389 children attending Head Start during the summer of 1965 and again in 1969 with 1531 children from over 150 Head Start classes across the nation. The sample children ranged in age from 3-0 to 6-5; 68.2 percent were Black, 15.9 were Mexican-American, 16.5 percent were White, 5.1 percent were Polynesian, and 4.2 percent were other (Puerto Ricans, Orientals, American Indians, and Eskimos).

Practical Considerations:

Compatibility:

The PSI test items and norming sample are very congruent with the types of children served in Chapter 1 preschool programs.

Administration:

1. The PSI takes less than 15 minutes to complete.

2. The PSI is administered individually by an examiner. Cues for what the examiner is to say to a child are clearly specified and guidelines are provided for scoring responses. The child's required responses may be oral, pointing, or motor, as appropriate.

3. The examiner may be a trained paraprofessional.

Scoring:

1. All items are scored as either correct (1 point) or incorrect (0 points). No distinction is made between a wrong answer and no answer (child silent or says he/she doesn't know). A child's score is the number of correct responses he/she makes. The maximum possible on the most recent revision of the PSI is 32.

Language Fairness:

A Spanish version of the PSI - Revised is available.

Effectiveness:

The PSI-Revised has been used in numerous large scale research studies that explored the effectiveness of preschool programs. These include:

- the 1968-69 Head Start National Study conducted by RTI;
- the 1966-72 Head Start Longitudinal Study (ETS);
- the 1969-71 Head Start Planned Variations Project (SRI, Huron Institute);
- a 1971 Project Follow Through pilot project;
- two Home Start Evaluations conducted through 1980 (High/Scope);
- the National Day Care Study and the National Day Care Home Studies conducted in 1975-81 (Abt Associates);

- the 1979-83 Child and Family Resource Program Evaluation (Abt Associates);
- the 1986-87 Home-Based Head Start Evaluation (RMC Research);
- the ongoing Project Giant Step Evaluation on New York City (Abt Associates).

The PSI has consistently yielded significant results in terms of magnitude of PSI change. Reliability measures reported by Abt Associates included a pre-posttest correlation of .67.

APPENDIX D

CHAPTER 1 PRESCHOOL PROGRAM OBJECTIVES, INSTRUCTIONAL APPROACHES, AND TESTING PRACTICES

From Telephone Interviews Conducted in January and February 1990

TABLE D.1: URBAN CHAPTER 1 PRESCHOOL PROGRAMS -- BACKGROUND INFORMATION

Questions		U1	U2	U3	U4	U5
1.	Number of Chapter 1 preschool classrooms	35	35	?	13	6
2.	Number of children enrolled per classroom	18	20	13	15	17
3.	Are children all Chapter 1 eligible	YES	YES	YES	YES	YES
4.	Do Chapter 1 classrooms follow school calendar	YES	YES	YES	YES	YES
5.	Last day of classes in the spring	6/1	6/1	5/30	6/1	6/1
6a.	Child testing done in Chapter 1 preschool	YES	YES	YES	YES	YES
6b.	Child testing done in kindergarten	NO	NO	NO	NO	YES
6c.	LEP children assessed with particular instruments	NO	NO	YES	NO	NO
7a.	Tests used	▪ DIAL-R	▪ Preschool Lang. Scale	▪ Penn. Preschool Inventory	▪ Denver Developmental ▪ Kindergarten Inventory of Dev. Skills	▪ Peabody Picture Vocab. Test-Revised ▪ Dallas Preschool Inventory
7b.	Tests used with LEP children	--	--	▪ Spanish version of PPVT-R	--	--
8.	Testing cycle	Pre: 6/1 Post: 5/1	Pre: 9/1 Post: 5/1	Pre: fall Post: spring	Pre: summer Post: spring	Pre: fall Post: spring
9.	Will send written program description	YES	YES	NO	YES	NO
10.	Objectives of the Chapter 1 preschool program	▪ Language enrichment ▪ Parent Involvement ▪ Self-esteem ▪ Basic skills	▪ Language enrichment	▪ Unit base incorporating the whole child; both experiential and concrete	▪ Child and parent ▪ Cognitive development	▪ Academic readiness ▪ Self-concept ▪ Peer socialization
11.	Number of years child may attend Chapter 1 preschool	1 year	1 year	1 year	3 years	1 year
12.	Program options after Chapter 1 preschool	Kindergarten	Kindergarten	Kindergarten	Kindergarten	Kindergarten
13.	Subsequent kindergarten enrollment options	75% same building	Same building	Same building	Same building	One third in same building
14.	Key differences between Chapter 1 preschool and K program	Pre-K: Developmental K: Academic	Pre-K: Experiential K: Teacher dominated	No differences	Pre-K: Activity K: Academic	Pre-K: Developmental K: Academic
15.	Estimated time to obtain parental consents	1 week	Recommend face-to-face communication and not mailing forms	2 weeks	2 weeks	2 weeks
16.	Chapter 1 program collects family background info.	Some info.	NO	Some info.	Some info.	Some info.
17.	Types of family background info. collected	Not specified	NONE	▪ Free lunch application	Not specified	Not specified
17a.	Will send copy of form used to collect family background info.	YES	--	YES	NO	NO

TABLE D.2: RURAL CHAPTER 1 PRESCHOOL PROGRAMS -- BACKGROUND INFORMATION

Questions	R1	R2	R3
1. Number of Chapter 1 preschool classrooms	5	4	1
2. Number of children enrolled per classroom	8	12	18-20
3. Are children all Chapter 1 eligible	NO	YES	NO
4. Do Chapter 1 classrooms follow school calendar	YES	YES	YES
5. Last day of classes in the spring	6/15	6/1	5/15
6a. Child testing done in Chapter 1 preschool	YES	YES	YES
6b. Child testing done in kindergarten	NO	NO	NO
6c. LEP children assessed with particular instruments	NO	NO	NO
7a. Tests used	▪ Early recognition intervention network (ERIN)	▪ Preschool Inventory	▪ Developmental test of Kindergarten readiness ▪ Golman, et al. Articulation Test ▪ Otis-Belensky motor proficiency
7b. Tests used with LEP children	--	--	▪ Vision, hearing and health exams
8. Testing cycle	Pre: Sept 15-30 Post: May 15-30	Pre: Screen 5/15	Pre: fall or late summer
9. Will send written program description	YES	YES	YES
10. Objectives of the Chapter 1 preschool program	▪ Whole language ▪ School/Kindergarten readiness	▪ Language development ▪ Pre-readiness	▪ Compensatory readiness for developmentally delayed
11. Number of years child may attend Chapter 1 preschool	1	1	1
12. Program options after Chapter 1 preschool	Kindergarten	Kindergarten	Kindergarten
13. Subsequent kindergarten enrollment options	Same building	Same building	Same building
14. Key differences between Chapter 1 preschool and K program	Programs closely coordinated; whole language emphasis is district wide	Pre-K and K are closely coordinated	Pre-K: Readiness K: Academic
15. Estimated time to obtain parental consents	2 weeks	2 weeks	2 weeks
16. Chapter 1 program collects family background info.	Incomplete	Incomplete	Some info.
17. Types of family background info. collected	▪ Free lunch and food stamp forms	▪ Free lunch and food stamp forms	--
17a. Will send copy of form used to collect family background info.	YES	YES	YES

TABLE D.3: URBAN/RURAL AND SUBURBAN CHAPTER 1 PRESCHOOL PROGRAMS – BACKGROUND INFORMATION

Questions	U-R1	S
1. Number of Chapter 1 preschool classrooms	3	47
2. Number of children enrolled per classroom	16	18-20
3. Are children all Chapter 1 eligible	YES	MIX
4. Do Chapter 1 classrooms follow school calendar	YES	YES
5. Last day of classes in the spring	6/15	6/15
6a. Child testing done in Chapter 1 preschool	YES	YES
6b. Child testing done in kindergarten	NO	NO
6c. LEP children assessed with particular instruments	NO	NO
7a. Tests used	▪ Syracuse Development Screening	▪ Language Section of Bohem-Slater
7b. Tests used with LEP children	--	--
8. Testing cycle	Pre: fall Post: spring	Pre: Sept. Post: May
9. Will send written program description	NO	NO
10. Objectives of the Chapter 1 preschool program	▪ Developmental	▪ Language development
11. Number of years child may attend Chapter 1 preschool	1 year	1 year
12. Program options after Chapter 1 preschool	Kindergarten	Kindergarten
13. Subsequent kindergarten enrollment options	Same building	Same building
14. Key differences between Chapter 1 preschool and K program	No differences	No differences
15. Estimated time to obtain parental consents	2 weeks	2-3 weeks
16. Chapter 1 program collects family background info.	YES	NO
17. Types of family background info. collected	▪ Use a common form for all preschool programs in county: very comprehensive	N/A
17a. Will send copy of form used to collect family background info.	YES	NO